



**Hugo Filipe Rangel da
Costa**

**Lanthipeptides of Archaea: the case study of
*Haloferax mediterranei***

**Lantipéptidos em Archaea: o caso de estudo de
*Haloferax mediterranei***

DECLARAÇÃO

Declaro que este relatório é integralmente da minha autoria, estando devidamente referenciadas as fontes e obras consultadas, bem como identificadas de modo claro as citações dessas obras. Não contém, por isso, qualquer tipo de plágio quer de textos publicados, qualquer que seja o meio dessa publicação, incluindo meios eletrônicos, quer de trabalhos acadêmicos.



**Hugo Filipe Rangel da
Costa**

**Lanthipeptides of Archaea: the case study of
*Haloferax mediterranei***

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biologia Molecular e Celular, realizada sob a orientação científica de Tânia Isabel Sousa Caetano, Investigadora em Pós-Doutoramento no Departamento de Biologia da Universidade de Aveiro e Sónia Alexandra Leite Velho Mendo Barroso, Professora Auxiliar com Agregação do Departamento de Biologia da Universidade de Aveiro.

Dedico este trabalho aos meus pais pelo incansável apoio e interminável paciência.

o júri

presidente

Prof. Doutora Maria de Lourdes Gomes Pereira
professora associada com agregação da Universidade de Aveiro

Doutor Hugo Alexandre Carvalho Pinheiro Osório
investigador auxiliar do Instituto de Patologia e Imunologia Molecular da Universidade do Porto
(ipatimup)

Doutora Tânia Isabel Sousa Caetano
professora auxiliar convidada da Universidade de Aveiro

agradecimentos

Quero agradecer à Professora Sónia pela imediata aceitação do que era ao início do ano um completo estranho no seu laboratório. Mais importante ainda, as várias oportunidades criadas pela Professora para uma fácil integração no grupo.

Quero agradecer também à Doutora Tânia pela orientação e constante disponibilidade para informar e ajudar o que era no início do ano um leigo nas questões práticas de um laboratório de microbiologia. Agradeço também pela paciência que isso deve ter requerido até à conclusão deste trabalho.

Aos restantes membros do laboratório Joana Lourenço, Joana Barbosa, Cláudia e aos meus colegas de mestrado Beatriz, Sofia e Teresa agradeço pela disponibilidade diária para ajudar com as mais variadas questões mas mais importante pelas conversas, horas de almoço e companhia que tornaram toda a experiência muito mais agradável.

Finalmente, quero agradecer o esforço heróico dos meus pais ao longo dos anos para garantir uma educação de qualidade a este filho.

palavras-chave

RiPPs, lanthipeptidos, Archaea, *Haloferax mediterranei*, *Haloferax volcanii*, co-expressão, *Escherichia coli*

resumo

Os produtos naturais são ainda hoje a maior fonte de descoberta de novas moléculas. Os lantipeptidos são um grupo de produtos naturais que pertencem à família de péptidos sintetizados pelos ribossomas e sujeitos a modificações pós-traducionais (RiPPs). Estas modificações incluem a desidratação de Ser e Thr e a ciclização entre estes aminoácidos desidratados e a Cys, dando origem aos característicos anéis de lantionina (Lan) e/ou metilantionina (MeLan). Estas reações são catalisadas por diferentes tipos de enzimas (LanB, LanM, LanKC e LanL), que constituem a base da classificação destes péptidos. Estas proteínas são codificadas em várias espécies de bactérias e possuem homólogos tanto no domínio Eukarya como no domínio Archaea. Os objectivos deste trabalho foram: i) identificar os clusters de lantipeptidos presentes no genomas de Archaea, recorrendo a bases de dados públicas, ii) caracterizar os clusters identificados e, por último, iii) identificar os padrões de desidratação e ciclização dos lantipeptidos encontrados em *Haloferax mediterranei* ATCC 33500.

Em Archaea foram detectados apenas genes *lanM*, que estão confinados aos genomas de organismos pertencentes à classe Halobacteria. Uma procura “manual” in silico na região genómica de cada *lanM* permitiu identificar a presença de vários genes *lanA*. A análise da sequência dos lantipeptidos codificados por estes *lanAs* não identificou motivos conservados, mas permitiu a sua divisão em grupos diferentes. Os clusters que contêm estes genes também codificam proteínas de função desconhecida e transportadores ABC. Nenhum destes transportadores é do tipo SunT (LanT) que estão normalmente associados ao processamento de lantipeptidos modificados por LanMs. *Haloferax mediterranei* ATCC 33500 foi selecionado para uma caracterização mais aprofundada uma vez que possui três genes *lanM*: um no cromossoma (*medM1*) e dois num plasmídeo (*medM2* e *medM3*). Foram construídos três vetores que permitiram a co-expressão dos genes *medM* e His₆-*medA* em *Escherichia coli*. Após produção, purificação por IMAC e análise por MALDI-TOF, verificou-se que nenhum dos péptidos foi desidratado. Desta forma, as enzimas MedM não parecem estar funcionais no citoplasma de *E. coli*, necessitando provavelmente de elevadas concentrações de sal para exercer as suas funções. Desta forma, utilizou-se um outro hospedeiro para expressão heteróloga: a Archaea *Haloferax volcanii* H1424. Para tal, foram construídos novos vectores de co-expressão baseados no plasmídeo pTA1392, contendo dois promotores (e não apenas um como na versão original). Após transformação de *H. volcanii*, os péptidos foram produzidos e purificados por IMAC. A análise por MALDI-TOF, não revelou a presença de nenhum dos His₆-MedA. Consequentemente, será necessário otimizar os protocolos de produção e/ou purificação destes péptidos de *H. volcanii*. O sucesso desta optimização permitirá provar a funcionalidade das enzimas MedM e estabelecer um sistema de expressão de natureza halófila que poderá ser útil para o estudo de outros clusters biosintéticos de Archaea.

Concluindo, organismos do domínio Archaea, em particular Halobacteria, codificam enzimas envolvidas na biossíntese de lantipeptidos homólogos às encontradas em bactérias. Os lantipeptidos de bactérias normalmente são bacteriocinas. Se os lantipeptidos de Archaea possuírem atividade contra outras Archaea (archaeocinas), poderão ser uma ferramenta valiosa para o desenvolvimento de novos marcadores seletivos. Isto permitirá melhorar as ferramentas existentes para manipulação genética de Archaea.

keywords

RiPPs, lanthipeptides, Archaea, *Haloferax mediterranei*, *Haloferax volcanii*, co-expression, *Escherichia coli*

abstract

Natural products are to this day the major source of novel molecules. Lanthipeptides are a group of natural products belonging to the family of ribosomally synthesized and post-translationally modified peptides (RiPPs). Their modifications include the dehydration of Ser and Thr residues immediately followed by a cyclization reaction between the dehydrated amino acids and Cys to form their characteristic lanthionine (Lan) and/or methyllanthionine (MeLan) rings. These reactions are catalyzed by different types of enzymes (LanB, LanM, LanKC and LanL), that define the classification of lanthipeptides. These proteins are encoded in a wide range of bacterial species and have homologues also in Eukarya and Archaea.

The objectives of this work were: i) identify lanthipeptide clusters in the publicly available archaeal genomes, ii) characterize the archaeal lanthipeptide gene clusters and iii) identify the dehydration and cyclization pattern of putative lanthipeptides from *Haloferax mediterranei* ATCC 33500.

Archaea encodes exclusively LanM enzymes (21 *lanM* genes) and they are confined to organisms of the class Halobacteria. Manual search of the genetic neighborhood of each *lanM* allowed to identify several *lanA* genes. The putative lanthipeptides encoded by these *lanA*s do not share conserved motifs but, based on their sequences, they can be divided into groups. Their gene clusters also encode proteins of unknown function and ABC transporters. None of these transporters are of the SunT-type (LanT) that is classically involved in the processing of LanM-modified lanthipeptides. *Haloferax mediterranei* ATCC 33500 was selected for further characterization because it has three *lanM* genes: one on the chromosome (*medM1*) and two on a plasmid (*medM2* e *medM3*). Three expression vectors were constructed to allow the co-expression of *medM* and His₆-*medA* genes in *Escherichia coli*. After production, purification by IMAC and MALDI-ToF analysis, it was found that none of the peptides were dehydrated. This indicates that MedM enzymes are not functional in *E. coli* cytoplasm. Most probably, their functionality is dependent on high concentrations of salt. Therefore, another heterologous expression host was used: the archaeon *Haloferax volcanii* H1424. To this purpose, new co-expression vectors were constructed based on the pTA1392 plasmid and containing two promoters, instead of the original one. After transformation of *H. volcanii*, peptide production, IMAC purification and MALDI-ToF analysis, none of the His₆-MedA peptides were identified. Thus, in the future, optimization of the production procedures and/or purification of these peptides from *H. volcanii* will be needed. This will prove the functionality of MedM enzymes and will establish a system with halophilic nature that can be useful for the investigation of other biosynthetic clusters from Archaea.

In conclusion, Archaea, in particular Halobacteria, encode lanthipeptide enzymes that are homologous to those found in Bacteria. In Bacteria, these peptides commonly have antibacterial activity. If archaeal lanthipeptides are archaeocins, these peptides will be a valuable tool for the development of novel selective markers for Archaea. This will be highly relevant to improve the toolbox for the genetic manipulation of Archaea.

Contents

Contents	i
List of figures	vii
List of tables	xiii
List of abbreviations	xv
1. Introduction	1
1.1 Natural products and RiPPs	3
1.2 Lanthipeptides	4
1.2.1 Lantibiotics – unique mode of action	4
1.2.2 The different classes of lanthipeptides	5
1.2.2.1 Class I Lanthipeptides	5
1.2.2.2 Class II Lanthipeptides	7
1.2.2.3 Class III e IV lanthipeptides	9
1.3 Leader peptide role and recognition in all classes of lanthipeptides	10
1.4 Additional PTMs overview	11
1.4.1 Protease and export activity	11
1.4.2 Tailoring enzymes	12
1.5 Immunity proteins	13
1.6 Regulation	14
1.7 Genome Mining	15
1.8 Lanthipeptides heterologous expression in <i>Escherichia coli</i>	16
1.9 The third domain of life – Archaea	16
1.10 <i>Haloferax mediterranei</i>	18
1.11 Work Objectives	19
2. Materials and Methods	21

2.1 Bioinformatics-----	23
2.2 Media and growth conditions-----	23
2.3 DNA extraction-----	24
2.4 <i>E.coli</i> co-expression system-----	24
2.4.1 <i>medA</i> and <i>medM</i> amplification and plasmid construction for <i>E. coli</i> expression -----	24
2.4.2 Co-expression and purification of the <i>medA</i> and <i>medM</i> products -----	25
2.5 <i>Haloferax volcanii</i> expression system -----	26
2.5.1 <i>medA</i> and <i>medM</i> amplification and plasmid construction for <i>Haloferax volcanii</i> co-expression-----	26
2.5.2 <i>Haloferax volcanii</i> H1424 transformation -----	27
2.5.3 MedAs expression in <i>Haloferax volcanii</i> H1424 -----	27
2.6 MALDI-TOF Mass spectrometry analysis -----	28
3. Results and discussion -----	29
3.1 Identification of CylM homologues in Archaeal genomes -----	31
3.2 Conserved residues between CylM and the putative LanM-----	34
3.3 Search and analysis of the putative archaeal precursor peptides -----	36
3.3.1 Identification of gene clusters encoding the biosynthesis of Class II lanthipeptides in Euryarchaeota-----	39
3.4 The case study of <i>Haloferax mediterranei</i> ATCC 33500 -----	42
3.4.1 <i>Escherichia coli</i> co-expression of the putative <i>medA</i> and <i>medM</i> genes from <i>Haloferax mediterranei</i> ATCC 33500 -----	44
3.4.2 <i>Haloferax Volcanii</i> H1424 co-expression of the putative <i>medA</i> and <i>medM</i> genes from <i>Haloferax mediterranei</i> ATCC 33500 -----	46
3.4.2.1 Plasmid pTA1392 construction for co-expression -----	46
3.4.2.2 <i>Haloferax volcanii</i> H1424 protein expression results -----	47
4. Final Remarks -----	51

5. Bibliography	55
6. Appendices	63
Appendix 1. Remainder of clusters identified through Bagel not used for comparison	65
Appendix 2. Genes present in chromosomal and plasmid clusters of <i>Haloferax mediterranei</i> ATCC 33500	66
Appendix 3. Predicted masses for His₆-tagged LanA peptides for <i>E. coli</i> co-expression study	67

List of figures

Figure 1 RiPPs biosynthesis schematic from Repka et al 2017 ⁷ . -----	3
Figure 2 A) The two step mechanism of lanthipeptide ring formation B) The four classes of enzymatic machinery responsible for the PTMs. Schematic from Zhang et al, 2015 ¹⁷ . -----	5
Figure 3 Dehydration step catalyzed by LanBs. In light grey are the conserved NisB residues thought to be involved in the reaction. Illustration from Repka et al, 2017 ⁷ . -----	6
Figure 4 Proposed cyclization mechanism, using nisin ring formation and NisC activity as an example. His212 might supply the proton for final ring structure, represented as H-A. Image adapted from Repka et al, 2017 ⁷ . -----	7
Figure 5 Dehydration through phosphorylation catalyzed by CylM. Adapted from Repka et al, 2017 ⁷ . -----	8
Figure 6 Mechanism for class III and IV dehydration through phosphorylation. In blue are the conserved residues of VenL. Image from Repka et al, 2017 ⁷ . -----	10
Figure 7 Gene cluster representation from class I and class II lanthipeptides. Example from Nisin and Lactacin 481. In green are the genes responsible for ring formation and lanthipeptide classification. In class I those are the lanB and lanC genes. In class II is the lanM gene. lanT and lanP genes are usually responsible for export and cleavage of the precursor peptide in class I. A bifunctional lanT assumes that role in class II. In red are the precursor peptides. In yellow immunity proteins. lanRK is the two component regulator of nisin. Image adapted from Alkhatib et al, 2012 ⁵⁰ . -----	16
Figure 8 Multiple sequence alignment of the identified LanMs. Analysis showed strong conservation of the residues involved in dehydration. -----	35
Figure 9 Multiple sequence alignment of the identified LanMs shows conservation of the three cysteine ligands for zinc. -----	35
Figure 10 Sequence similarity network of LanMs generated using EFI-EST and visualized in Cytoscape with an alignment score threshold of 110 (~35% sequence identity). -----	36
Figure 11 Multiple Sequence Alignment of all manually identified putative precursor peptides. In brackets are the four distinct groups identified. -----	37

Figure 12 Multiple sequence alignment between the four groups of similar putative LanA peptides. -----39

Figure 13 Comparison between the four distinct groups of putative gene clusters identified by Bagel. The ABC transporters are represented in red, and in green are represented the putative lanA genes identified by the software. In the *Haladaptatus paucihalophilus* DX253 cluster two lanthipeptide regulation related genes were identified, in yellow. Two other genes, in cyan, were also identified as associated with lanthipeptide clusters but with undefined function. The Tn represents genes encoding transposases. The black lines represent the minimal genes that should be involved in the biosynthesis of each lanthipeptide. -----41

Figure 14 Schematic representation of the two clusters identified in *Haloferax mediterranei*. In blue the LanMs genes identified. In yellow the putative peptides genes. In red the ABC transporter genes. -----43

Figure 15 Plot generated by TMHMM analysis of the gene HFX_RS16010 coding for the transmembrane domain of an ABC transporter. -----43

Figure 16 Sequence of the His₆-tagged precursor peptides for *E. coli* co-expression. Colored in red are the serine and threonine residues essential for the dehydration reaction. -----44

Figure17 SDS-Page of whole cell proteins. The cells harboring the gene medM3 were used for protein expression confirmation. Results from 4 and 18 hours incubation at 37°C and 18°C after induction of expression by IPTG confirmed expression of our 120 kDa enzyme (arrows). Respective negative controls are present. -----45

Figure 18 Schematics of plasmid construction approach used for co-expression of medA and medM genes in *Haloferax volcanii*. a) vector pTA1392 with medA gene, b) vector pTA1392 with medM gene, *) sequence of the promoter and the medA gene fused to the His₆-Tag, c) final construct containing both genes associated with their p.tnaA promoters. -----47

Figure 19 Sequence of the His₆-tagged precursor peptides for *H. volcanii* co-expression. Colored in red are the serine and threonine residues essential for the dehydration reaction. -----48

Figure 20 Mass spectra of *Haloferax volcanii* expression products after IMAC purification and ZipTip desalting. A) *Haloferax volcanii* transformed with plasmid containing the chromosomal medA1; B) *Haloferax*

volcanii transformed with plasmid containing *medA2*; C) *Haloferax voltanii* transformed with plasmid containing *medA3*.-----48

Figure 21 Putative lanthipeptide clusters identified through Bagel. In blue are the genes encoding the modifying enzymes *LanM* and in red the ABC transporter related gene. In green are the putative peptides identified by the software. Curiously, in the *Halopiger djelfimassiliensis* cluster the software identifies the domain DUF4135 (*LanM* dehydration domain) and the *LanC* domains as separate proteins. Homology between *Natrinema gari* JCM 14663 and *Natrinema* sp. J7-2 might indicate that we are in the presence of the same organism. -----65

List of tables

Table 1 Primers used for gene amplification by PCR for *E. coli* expression. The enzymes shown were the enzymes used for cloning and the *Ta* represents the temperature of annealing used in each PCR reaction. -24

Table 2 PCR reaction used for amplification of genes. -----24

Table 3 PCR conditions used for gene amplification. * represents the *Ta* used in each reaction and that is shown in Table 1. -----25

Table 4 Primers used for gene amplification by PCR for *Haloferax volcanii* expression. -----26

Table 5 Summary of all strains obtained from transformation. -----27

Table 6 CylM homologues identified by PSI-Blast. All organisms belong to phylum Euryarchaeota, class Halobacteria. -----32

Table 7 Genes present in chromosomal and plasmid clusters of *Haloferax mediterranei* ATCC 33500 -----66

Table 8 Predicted masses for His₆-tagged LanA peptides for *E. coli* co-expression study. Each serine or threonine residue belonging to the core sequence can be subject to a dehydration reaction. Given that we have no means of ascertaining where leader sequence ends and core sequence begins an overestimation of the number of reactions is needed. For the loss of each water molecule, a shift of ~18Da is expected. -----67

List of abbreviations

ATP Adenosine triphosphate

ABC protein ATP binding cassette protein

Dha Dehydroalanine

Dhb Dehydrobutyrine

Dhx Dehydrated residues

FNLD PheAsnLeuAsp motif

IMAC Immobilized metal ion affinity Chromatography

Lab Labionin rings

Lan Lanthionine rings

LanB Class I dehydratase protein

LanC Class I cyclase protein

LanFEG ABC transport complex involved in lanthipeptide selfimmunity

LanI Immunity protein

Lan KC Class III modifying enzyme

LanL Class IV modifying enzyme

LanM Class II modifying enzyme

LanP Protease enzyme

LanT Transport enzyme

MFS Major Facilitator Superfamily

MeLan Methyllanthionine

MSA Multiple Sequence Alignment

MCS Multi cloning sites

NRPSs Multimodular nonribosomal peptide synthetases

ORFs Open reading frames

(PSI)-BLAST Position-Specific Iterative Blast

PSSM Position-specific score matrix

PTMs Post translational modifications

RiPPs Ribosomally synthesized and post-translationally modified peptides

SSN Sequence Similarity Network

1. Introduction

1.1 Natural products and RiPPs

Molecular evolution on Earth is responsible for the extraordinary library of chemical structures present in millions of species and impossible to reproduce by human hand either in number, diversity and function. Besides their natural role in complex biological systems, these molecules often have a lot to offer to the scientific and medical communities. In fact, many of the major therapeutic drugs commercially available from the past 50 years are natural products or based on their chemical skeleton¹. The classical ways of obtaining natural products either by extraction or organic synthesis are now being complemented by the massive amounts of genetic information being amassed from all domains of life and the development of systems that allow the manipulation of biosynthetic pathways and the discovery of novel molecules².

Ribossomally synthesized and post-translationally modified peptides (RiPPs) are a large class of structurally diverse natural products that is gaining considerable attention for their biological activities and characteristic biosynthetic pathways. Capable of a similar degree of chemical diversity thought possible only when referring to nonribosomal peptides, through the activity of the large, multimodular nonribosomal peptide synthetases (NRPSs), RiPPs rely instead on extensive post translational modifications (PTMs) of a ribosomally synthesized precursor peptide³. These modifications are responsible for the variety of chemical structures and consequently the wide range of activities, from antibiotics⁴ and antivirals⁵ to allelopathics⁶.

RiPPs are divided into different subfamilies based on their characteristic biosynthetic machinery and structural features. Despite subfamily, RiPPs biosynthesis starts with the ribosomal translation of the precursor peptide. This precursor peptide usually contains an N-terminal leader sequence and a C-terminal core region that will be subject to the post translational modifications. The biosynthetic machinery then recognizes the leader sequence (sometimes a signal sequence is also present in the C-terminal region) and catalyzes the various amount of PTMs. Thereafter, the leader sequence of the peptide is cleaved by peptidases to yield the mature product. Removal of the leader sequence usually translates into an active peptide (Figure 1). Lanthipeptides are a RiPP subfamily which are gaining considerable interest.

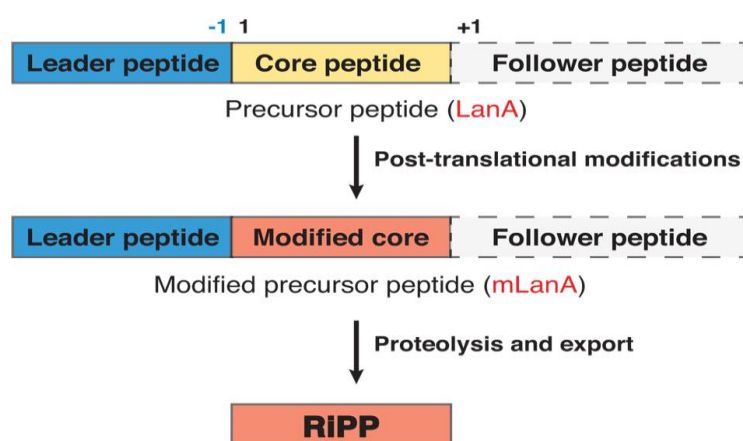


Figure 1 RiPPs biosynthesis schematic from Repka et al 2017⁷.

1.2 Lanthipeptides

Lanthionine containing peptides, or lanthipeptides, are a class of molecules that contain in their structure lanthionine (Lan) and methyllanthionine (MeLan) residues. They are the largest group of RiPPs based on the frequency of their biosynthetic gene clusters in currently available genomes⁸. These residues are modified in a characteristic two-step mechanism. First the serine and threonine amino acids present in the precursor peptide are dehydrated to dehydroalanine (Dha) and dehydrobutyrine (Dhb), respectively. Then follows cyclization where the thioether crosslinks are formed via a Michael-type like reaction done by a nucleophilic attack of Cysteine to the dehydro amino acids. The formed rings are what give these peptides their stable conformational structure and are also essential for their biological activities. The present classification system allows for four different classes of lanthipeptides depending on the biosynthetic enzymes that catalyze these reactions⁹. A recent attempt for a common nomenclature determined that the biosynthetic genes for lanthipeptides attributed the generic symbol *lan*, with a more specific designation for each lanthipeptide molecule⁹. For example the gene coding for the precursor peptide of nisin is termed *nisA* and for its biosynthetic enzymes genes *nisB* and *nisC*.

Dating back to 1928, Nisin, produced by *Lactococcus lactis*, was the first reported lanthipeptide. Given their complex and unusual composition of amino acids, lanthipeptides were thought to be the product of a non-ribosomal mechanism. It wasn't until the 1980's that the ribosomal origin of this class of peptides was verified in the laboratory, with the identification of the gene for the precursor peptide of epidermin¹⁰. Genes encoding homologues of lanthipeptide enzymes seem to be present in all domains of life. For instance, a LanC like protein was attributed the role of a transcription suppressor when overexpressed in human tumors¹¹ and that of a kinase regulator in human liver cells¹², but peptide isolation is so far restricted to bacteria.

1.2.1 Lantibiotics – unique mode of action

Nisin brought special attention to this class of RiPPs due to its long standing action as an antimicrobial in the food industry without substantial incidence of microbial resistance¹³. The lanthipeptides with antimicrobial activity are called lantibiotics. Isolated lantibiotics possess potent activity against many of the today clinically relevant strains of both the Gram-positive and some of the Gram-negative bacteria. Many lantibiotics display highly specific activity against a rather limited spectrum of sensitive strains and a lower activity against a broader range of organisms.⁴ This high activity and low incidence of resistance can be explained by the dual mode of action of these peptides. Molecules like nisin, subtilin and epidermin have the ability to both affect cell wall biosynthesis and disrupt the cytoplasmatic membrane integrity by pore formation¹⁴. The formed pores allow the efflux of ions and small molecules and the inevitable dissipation of the membrane potential. The same authors that demonstrated the hydrophilic conformation of nisin in water postulated that the molecule may have a different behavior in its target site, the phospholipid membrane. In a conformational study using membrane mimetics they concluded that the molecule adopts an amphipathic conformation that enables the peptides to reach into the bacterial membrane, inducing enough perturbations to the phospholipids that allow their eventual assembly into a pore¹⁵. It is now known that nisin uses lipid II, the central building block of the membrane wall, as a docking molecule for pore formation. The two N-terminal rings of nisin form a stable complex, called the pyrophosphate cage that envelops the

bactoprenol moieties of the lipid II intermediate of cell wall biosynthesis. This binding pocket is stabilized by hydrogen bonds involving several residues of nisin. After binding to lipid II, the C-terminus of nisin is able to insert into the membrane, oligomerize and form a pore that contains eight nisin molecules and four lipid II molecules¹⁶. Pore formation might also be the reason why nisin is capable of inhibiting the growth of bacterial spores. Although this might not be true to all lantibiotics, the conservation of a nisin like structure possibly endows these molecules with a similar mode of action.

1.2.2 The different classes of lanthipeptides

The most recent and broadly used classification system for this class of RiPPs is based on the catalytic machinery responsible for the dehydration and cyclization reactions⁹ (Figure 2). Presently this family of RiPPs is divided in four classes, discussed with further detail below.

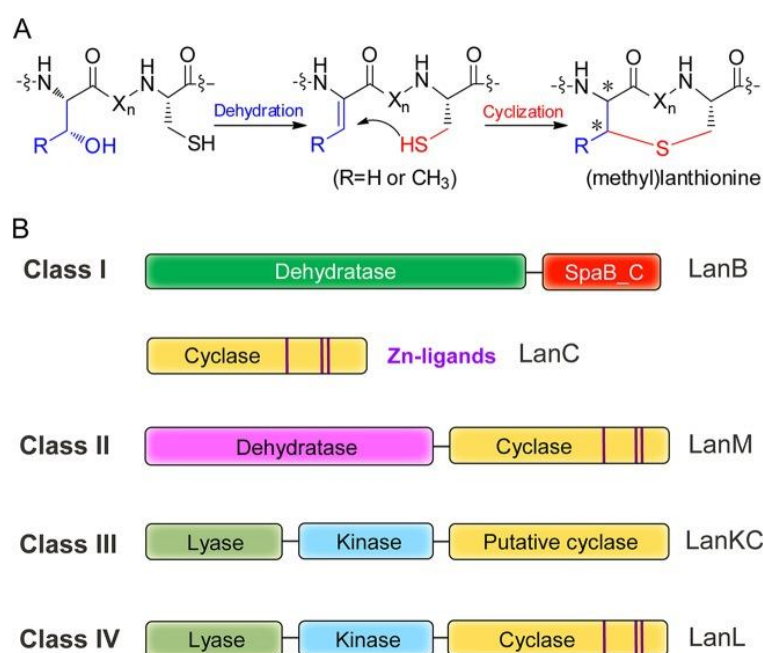


Figure 2 A) The two step mechanism of lanthipeptide ring formation B) The four classes of enzymatic machinery responsible for the PTMs. Schematic from Zhang et al, 2015¹⁷.

1.2.2.1 Class I Lanthipeptides

The prototypical lanthipeptide nisin belongs to class I lanthipeptides. The biosynthesis of nisin is encoded by a transcriptional operon of 11 genes. This operon encodes the precursor peptide, NisA and several others proteins responsible for the characteristic lanthipeptide PTMs, precursor peptide cleavage, transport and also immunity from the mature peptide. The dehydratase NisB catalyzes the dehydration of serine and threonine in the core peptide. The cyclase NisC catalyzes the attack of cysteine to the dehydrated residues generating the five lanthionine rings of nisin. NisP, a protease, removes the leader sequence from the precursor peptide, allowing for the final, mature, bioactive product. To avoid self-destruction, genes for immunity or self-preservation proteins, *nisI* and *nisFEG*, are also present. NisI is a lipoprotein that sequesters nisin. It was shown to act bound to the cell membrane but it could also be excreted to the medium as a potential

enhancer¹⁸. The ABC transporter homologous system, NisFEG, seems to work by expelling nisin molecules into the environment in a NisI independent fashion¹⁹. As they serve as a basis for lanthipeptide classification, an exploration of the mechanisms of the dehydration and cyclization will be further explored.

Dehydration by LanBs. Although nisin being known since the 30's, the knowledge of their ribosomally synthesized nature and the exact function of their biosynthetic enzymes came later on. Works like the one carried by Koponen et al (2002) made obvious the role of the genes *nisB* and *nisC* in these PTM²⁰. The expression of a histidine tagged *nisA* in *L. lactis* mutants, revealed the dehydratase function of NisB and the cyclase one of NisC. Although *in vivo* activity studies were yielding exciting results, *in vitro* studies revealed a challenge limiting our understanding of these modifications. It wasn't until 2013 that a peak at the reaction mechanics was possible. Glutamylation of serine and threonine residues in the precursor peptide was key for obtaining the dehydrated product (Figure 3). Mutational analysis revealed that the N-terminal domain of NisB was responsible for creating the glutamylated precursor intermediate, and posterior elimination of glutamate was attributed to a C-terminal domain²¹. These two domains seem to possess unique structure and they share no similarity with other known protein families. The current mechanistic models state that LanBs use the readily available activated glutamate present in the cellular medium (glutamyl-tRNA) to catalyze the dehydration reaction. A co-evolution event seems to be at play also. Differences in tRNA, namely the acceptor stem sequence determine the recognition of the glutamyl-tRNA by the dehydratase. This means that heterologous expression systems might not work in every case as an appropriate tRNA must be present²².

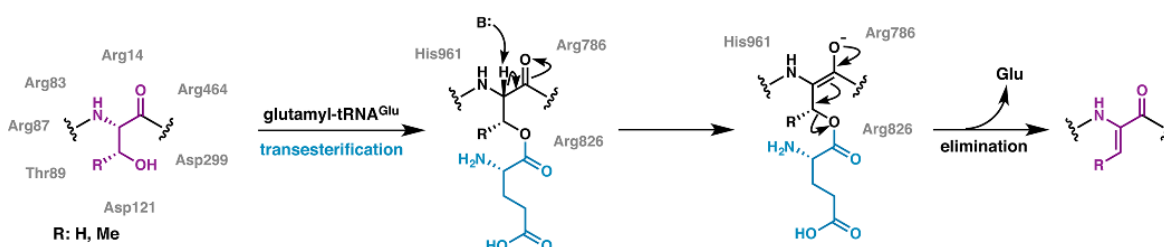


Figure 3 Dehydration step catalyzed by LanBs. In light grey are the conserved NisB residues thought to be involved in the reaction. Illustration from Repka et al, 2017⁷.

Cyclization by LanCs. The second step of lanthionine formation is done by the attack of the cysteine thiol group to the dehydrated residues (Figure 4). This reaction can be recreated in the laboratory, as it seems to not need the enzymatic action of the cyclase in a basic solution. Interestingly, cyclization occurrence does not mean correct overall peptide structure. Chemical synthesis studies with nisin revealed that the reaction products were a mix of different structures and NisC was needed for the assembly of the correct ring topology with the possible reason being that the reaction rate of Dha+Cys is higher than DhB+Cys⁷ resulting in abnormal ring formation. Nevertheless, other studies involving the elimination of the *lanC* genes suggested that cyclization will not occur in the cell cytoplasm without the enzyme. NisC is a metalloenzyme and crystallography data suggested that a zinc ion is complexed with the residues Cys284, Cys330, and His331. Mutational studies also attributed relevance towards correct cyclization activity to the NisC residues His212 and Asp141 that is hydrogen-bonded to His212 in the crystal structure²³. Conservation of these residues or chemically similar amino acids might be a good indication of

cyclase activity in putative novel enzymes identified in today's available genomes. The proposed model for this metalloenzyme reaction mechanism presumes that the zinc ion is essential for the deprotonation of the thiol group of cysteine by lowering its pKa as is observed in other similar enzymes. The sulfur is then ready for the nucleophilic attack on the dehydrated residue, Dha or Dhb, β - carbon. LanC enzymes seem also to be essential for correct ring stereoisomery. The characteristic stereochemistry of the MeLan ring in nisin seems possible only by the aid of a nearby amino acid. Is not yet understood how LanC enzymes control cyclization and determine a specific structure to a lanthipeptide as a number of combinations are possible between cysteine and the dehydrated residues to form a myriad of ring structures. Two hypothesis are in order. One way to reduce the number of possible combinations is having less dehydrated residues per cysteine at the time of cyclization. This means that although they act as standalone enzymes, LanB and LanC coupled activity seems to be essential for the correct structure of a number of lanthipeptides. Cyclization of a readily available dehydrated amino acid, determines where the next dehydration will occur, as a lanthipeptide ring might prevent a dehydration reaction on a nearby serine or threonine residue. Another possibility is that the substrate sequence has an inherent preference for a determined ring topology limiting LanC cyclization range. It is also quite possible that a mix of this strategies is used for the production of the final lanthipeptide structure

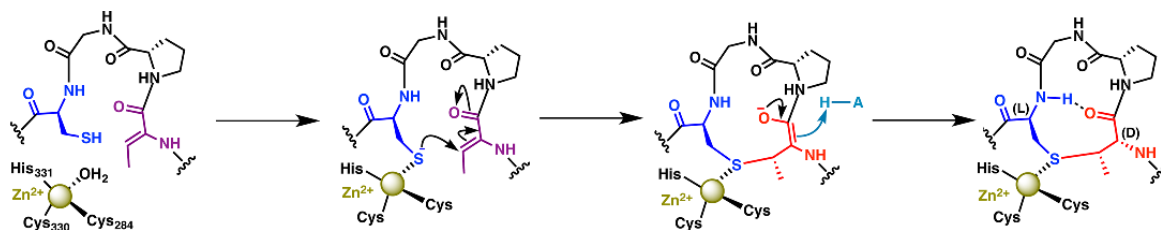


Figure 4 Proposed cyclization mechanism, using nisin ring formation and NisC activity as an example. His212 might supply the proton for final ring structure, represented as H-A. Image adapted from Repka et al, 2017⁷.

1.2.2.2 Class II Lanthipeptides

This class of peptides is characterized by the presence of bifunctional enzymes, LanMs, which are responsible for the dehydration and cyclization steps towards lanthine ring formation. A LanC like domain in the C-terminal region of the protein was immediately identified as the cyclization domain. The N-terminal region responsible for the dehydration step, had no homology with any known lanthipeptide domains. That is explained by the fact that, unlike class I lanthipeptides, the dehydration reaction is accomplished through phosphorylation instead of glutamylation⁸.

One subgroup of class II lanthipeptides is the two-component lantibiotics. These lanthipeptides display maximum activity when two different post-translationally modified peptides are used together. For most two-component lantibiotics, two different LanM enzymes carry out the dehydrations and cyclizations steps of the two different LanA substrates⁷. The exception is cytolysin, where only one LanM, CylM, is responsible for the PTMs. Cytolysin seems to lack antibiotic activity but instead acts as a virulence factor, lysing mammalian cells²⁴.

LanM N-terminal dehydration domain. A structural study of dehydratase domain of CylM, the first of its kind on LanMs, revealed that it resembles the catalytic core of eukaryotic lipid kinases,

despite the absence of clear sequence homology²⁵. Compared with canonical lipid kinases the CylM dehydration domain is significantly larger and contains additional subdomains. These are the kinase-activation domain, the P-loop and the C-lobe that seem to play a role in the reaction mechanism. Once again, mutational studies shed light on the importance of several residues in the dehydration step (Figure 5). Asp252 and His254, located on the P-loop appear to be involved in the activation of adenosine triphosphate (ATP) for serine/threonine phosphorylation. Asp347 and His349, located on the C-lobe accept a proton from serine/threonine for phosphorylation to occur. Lys274 also seems to take part not only in the initial steps of phosphorylation but also phosphate elimination. Other residues essential for phosphate elimination are Arg506 and Thr512 of the Kinase-activation domain. Residues Asn352 and Asp 364 are also conserved and seem to interact with ATP through a Magnesium ion. Other residues have been identified for LanM activity based on additional mutagenesis studies of a different LanM enzyme, BovM. These include Tyr230, Glu266, Glu446, and Gln495 in BovM that correspond to Tyr330, Glu366, Glu555, and Gln611 in CylM⁷.

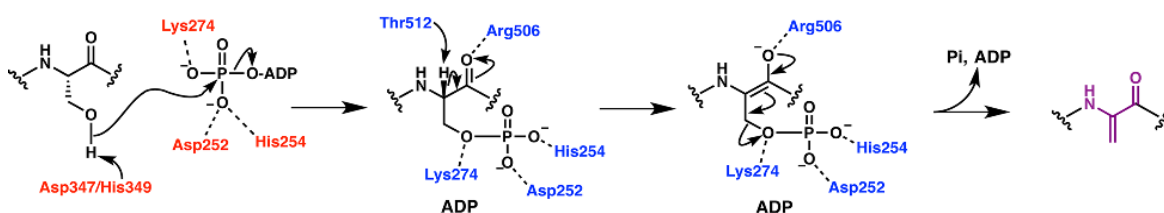


Figure 5 Dehydration through phosphorylation catalyzed by CylM. Adapted from Repka et al, 2017⁷.

LanM C-terminal cyclization domain. LanM C-terminal cyclization domain is analogous to LanC proteins, shares certain structural features and uses a metal ion cofactor indicating that the cyclization reactions should be highly similar (Figure 4).

The cyclase domain in CylM shares with NisC the histidine and two cysteine conserved residues that encapsulate the Zinc ion²⁵. In CylM these residues are Cys875, Cys911, and His912. Many LanMs contain three cysteine ligands for zinc instead of two cysteine ligands and the one histidine. LanM identified in Euryarchaeota are all part of this group⁷. The LanM cyclization domain also contains a conserved histidine residue as do LanC enzymes, suggesting a function for the residue. In NisC the histidine is located at position 212 in CylM in 790. The conservation of these residues in novel putative enzymes might indicate preservation of biological activity.

Enzymes are typically highly stereoselective catalysts that enforce a specific conformation on their substrates during the reaction. This usually means that a particular enzyme acts on specific substrates to obtain a particular reaction product. The cyclization events aided by some LanMs appear to differ from this norm. Some enzymes seem to be able to recognize a high number of different substrates and allow for a specific chemical structure for each and every one of the substrates. Other enzymes are responsible for the different stereoisomery in the same peptide which is highly intriguing given the existence of a single active site for the cyclization⁸. This raises the question – are ring topologies determined exclusively by the enzymes, by the substrates or by a combination of both? As stated above, Dha and Dhb non enzymatic reaction with cysteine seems to possess different reaction rates that might result in incorrect lanthipeptide ring topology. What also happens, in works where the sequence of the lanthipeptide was altered and subject to enzymatic activity to study the final products, is that the substrate controls the

stereoselectivity of the enzyme²⁶. What seems to be at play is a mixture of substrate and enzyme regulation of the final lanthipeptide structure. This type of control is to this day very difficult to replicate by chemists¹. This control of stereoisomery by LanM enzymes might have very interesting biotechnological applications someday.

1.2.2.3 Class III e IV lanthipeptides

These are the two most recently discovered classes of lanthipeptides and whose trifunctional enzymes (with a kinase, lyase and cyclase domains) share the same distinct dehydration mechanism. The cyclization domain in class IV enzymes is similar to the LanM and LanC but class III shares no homology with the rest^{27,28}. Class III enzymes receive the designation LanKC. Class IV, LanL⁸. Class III lanthipeptides have another interesting feature, the labionin rings (Lab)²⁹. These rings are formed through the incorporation of two different cross-links into the peptide. One is the carbacyclic ring in an N-terminal position and the other the thioether ring positioned in the C-terminal.

Dehydration via phosphorylation by two active sites. Dehydration of class III and IV lanthipeptides is achieved by two separate kinase and lyase domains present in each enzyme. Although no x-ray structures were determined at this point for these classes of biosynthetic proteins, *in vitro* and mutational studies revealed interesting information about these enzymes and their particular domains. First, unlike class II, there's no specific phosphate donor, meaning that there isn't a sole requirement for ATP. Different class III enzymes have different nucleoside triphosphate preferences. Some are even capable of using various nucleosides⁷. Having two active sites, one for phosphorylation of the substrate and another for phosphate elimination, means that the dehydration reaction is also not as linear as in class II. A study on the modification of the class III lanthipeptide curvopeptin by CurKC revealed that although the events seem to occur in a general C→N directionality, phosphorylation of different residues is not immediately followed by phosphate elimination³⁰. The modified substrate, either by phosphorylation or complete dehydration of particular residues determines which domain, kinase or lyase, will be acting. This interplay between active sites seems to play a relevant role for correct ring formation as incubation of a precursor peptide exclusively with the kinase domain of an enzyme, only then followed by incubation of the lyase domain resulted in dehydration inefficacy³⁰. Several mutational studies pinpointed the residues responsible for appropriate dehydration. Various residues seem conserved and with a role for phosphate elimination. In SpvC, a class III enzyme, Lys136 was pinpointed and in VenL Lys80. His106 in SpvC and correspondent conserved residue in VenL, His53, play a critical role also. Another lysine in SpvC, Lys104 with the corresponding residue Lys51 in VenL, are also relevant for proper function⁷.

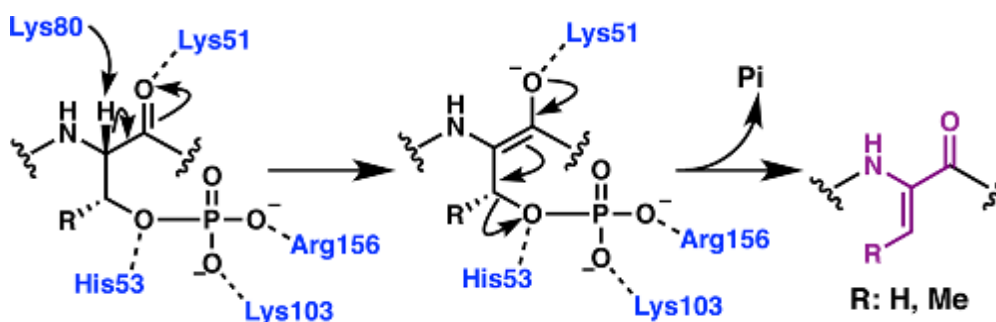


Figure 6 Mechanism for class III and IV dehydration through phosphorylation. In blue are the conserved residues of VenL. Image from Repka et al, 2017⁷

Cyclization domains in class III and IV. The remarkable characteristic of Class III lanthipeptides is the presence of not only lanthionine, but also labionin rings catalyzed by the cyclization domain of some LanKC, domain that shares no homology with any other class. Although some enzymes seem to favor the Lan formation, others appear to favor de labionin structure. Some enzymes appear to be capable catalyzing both formations in their respective peptides. Distinction between Lab and Lan rings residues is possible based on the reactivity of the dehydrated residues that did not cyclized. If a Lan structure is formed, Dhx react with thiols. If a Lab ring is preferred, Dhx lose their reactivity⁷. No proposed model of reaction mechanism is yet available capable of explaining this features.

Class IV LanL cyclase domain shares homology with the metalloenzymes of class I and II and share the same ligand site as some LanM enzymes associated with higher activity and ligand tolerance²⁸.

1.3 Leader peptide role and recognition in all classes of lanthipeptides

Although all PTMs being done to the core peptide and lanthipeptide bioactivity being linked to leader sequence removal, there's clear evidence for a role of this N-terminal region sequence in the enzymatic reactions and correct final product yield. The leader sequence interacts not only with the biosynthetic machinery but it may also interact with immunity and transport proteins in the producing organism.

Recognition by LanB and LanC enzymes. Several class I precursor peptides seem to share similar number and type of residues in the leader sequence. Mutational studies with NisA revealed a conserved motif of four amino acids, PheAsnLeuAsp (FNLD) involved in peptide biosynthesis. FNLD motif seems to be essential for LanB recognition while other residues of the leader sequence seem to play a role with cyclase LanC interaction. This FNLD-LanB interaction was visualized with a crystal structure of NisB in complex with its substrate peptide NisA²². The NisB residues Val176, Val198, Tyr202, Leu209 e Tyr213 are responsible for the interaction with phenylalanine of the FNLD motif. Il171, Tyr213 and Leu217 interact with the leucine residue. Finally, Arg154 with the asparagine one⁷. The FNLD motif is located in the N-terminal half of the leader peptide and the residues of the C-terminal region appear to be a spacer that enables the core peptide access to the catalytic site of the LanB responsible for glutamylation.

Recognition by LanM enzymes. CylM is the first enzyme to have a defined crystal structure and the domain responsible for leader sequence recognition in class I lanthipeptides seems not to be

preserved in class II. This is supported by the differences found in leader sequences of class II lanthipeptides. Instead of the characteristic motif FNLD found in class I, class II leader sequences are found to have a ELXXBX motif, with B being either hydrophobic residue, Valine, Leucine or Isoleucine. And instead of a N-terminal positioning, this motif is located in the C-terminal region of the leader sequence⁷.

Interestingly, some LanMs were shown to maintain activity even when the leader sequence of their original precursor peptides was fused with different core peptides. Although this is not always the case for every enzyme, there could be a biotechnological application for the very characteristic PTMs done by these proteins. The explanation for this promiscuity is that the leader sequence alone induces an active state of the LanM and facilitates precursor peptide modification³¹, even if not the original one. The current model for LanM catalysis dictates that leader sequence is responsible for the activation of the enzyme followed by some level of recognition of the core peptide for some enzymes, followed by modification. The core peptide sequence seems to also influence the order of dehydration reactions and final Lan ring structure⁷.

Recognition by LanKC and LanL enzymes. A series of mutational studies revealed the importance of the ILELQ motif present in the leader sequence for LanKC recognition³². As is the case in other classes, the residues present in the C-terminal region of the leader sequence are important to present the core peptide in the active site and allow catalysis. As recognition is done in a separate domain from the active site, a number of residues are necessary between the site of recognition and core peptide for dehydration to occur³².

Another motif was found in the leader sequence of RamS that contains an LFDLQ motif similar to the ILELQ one. This motif is widely found in class III precursor peptides⁷. This residue conservation might allow for the identification of new putative LanA precursor peptides in novel class III and class IV clusters identified by genome mining.

1.4 Additional PTMs overview

1.4.1 Protease and export activity

RiPPs and in this case, lanthipeptides, are usually not active until the leader peptide sequence is removed. This cleavage is usually coupled with mature peptide export from the cytoplasm.

In Class I. Analysis of class I gene clusters like the one belonging to nisin revealed the existence of a gene, LanT, whose sequence was homologous to membrane transporters. In fact, these transmembrane transporters have a conserved sequence, SpaT-like domain, that shares high similarities with ATP-binding proteins and are now known to be part of the ATP binding Cassette (ABC) transporter family. The export function of these proteins was confirmed with deletion of the gene in producer strains, resulting in loss of cell viability³³. Leader peptide sequence is essential for peptide transport. The core peptide, modified or not and even if belonging to a different precursor peptide presented to the LanT protein as a chimera, seems to have no influence in the mechanism of transport, as some substrate tolerance studies showed. What seems to also be necessary for a proper efflux of the peptide is an interaction between LanT, LanB and LanC meaning that these biosynthetic enzymes must also be associated to the cellular

membrane³⁴. As a general rule for class I lanthipeptides, following precursor peptide transport to the outside of the cell ensues cleavage of the leader sequence from the precursor peptide. This is achieved by a protease usually anchored to the cell wall. Sequence alignments of NisA with other nisin-like precursor peptides identified a conserved Gly-Ala-(Xxx)₂-Arg-Ile motif, where NisP cleaves between the Arginine and Isoleucine residues. The importance of this motif was confirmed once again with mutational studies, where alteration of some of the residues resulted in loss of protease activity and an immature peptide³⁵. More recently a second type of protease associated to the cytoplasm, not selective for the modified product and with a new leader peptide recognition motif was identified making more difficult for an easy identification of precursor peptides based on sequence homology. A fully modified core peptide seems to also be needed for cleavage to occur.

In Class II. In this class, the protease and transport function are all done by a bifunctional enzyme that shares the terminology LanT with class I. These proteins are members of the ABC-transporter maturation and secretion (AMS) protein family. Cleavage by the protease domain generally occurs after a two glycine conserved motif, the double-Gly motif. Glycine-Alanine and Glycine-Serine sequences are also observed variations of this motif. These enzymes seem to have broader substrate specificity.

A new series of secretion proteins were also identified in Gram-negative bacteria clusters. These proteins are part of the major facilitator superfamily (MFS) and are responsible for the efflux of the lanthipeptide from the cell⁷.

Besides the variation in cluster composition, more recently we are starting to see a blurring of the lines in class I and II distinction based on transport and cleavage function as bifunctional enzymes were identified in class I lanthipeptides and LanP genes seem regularly present in class II clusters⁷.

Class III and IV. The removal of the leader peptide of class III and IV lanthipeptides is very different from that of class I/II compounds. The most obvious hint is the lack of a gene encoding a dedicated protease in these classes' clusters. The evidence that the same precursor peptide can generate a number of different peptides with varying number of residues confirms the lack of activity of such protease. It appears that the class III and IV lanthipeptides maturation is done by a stepwise trimming of the leader peptides by currently a different class of peptidases³⁶.

1.4.2 Tailoring enzymes

Class I. Several additional post-translational modifications are present in class I lanthipeptides. Particular enzymes that are not necessarily present in the gene cluster introduce these modifications. One of such modifications is the addition of an N-terminal lactyl cap to the mature peptide. After cleavage of the leader sequence, the exposed N-terminal amino acid suffers a series of non-enzymatic modifications culminating in a final enzymatic reaction possible catalyzed by an oxidoreductase⁷.

Another modification, done this time to the C-terminal region of the peptide, consists of the decarboxylation of a cysteine residue. This modification was first identified in 1985 with the elucidation of the epidermin structure³⁷. A gene, *epiD* encoding for a protein belonging to the

homo-oligomeric Flavin containing Cys decarboxylase family was responsible for the modification³⁷. Several other peptides have the same modification and encode a similar protein in their genome.

Another set of modifications in class I lanthipeptides are tryptophan halogenation and proline hydroxylation. These PTM were identified when determining the structure and mode of action of microbisporicin³⁸.

Class II. This class of lanthipeptides seems to harbor a set of additional PTM of its own. One characteristic feature is the formation of the unusual D-Amino acids. The prevailing hypothesis states that after dehydration of serine, instead of ring formation, a different enzyme is responsible for the D-residue formation through hydrogenation. This protein may belong to two different families, the zinc-dependent alcohol dehydrogenases family or the Flavin-dependent oxidoreductases one⁷.

Cinnamycin, cinnamycin B, duramycin, duramycin B and duramycin C are an exclusive group of class II lanthipeptides with hydroxylated aspartate and lysinoalanine in their structure⁷.

Finally another unusual modification present in class II peptides is the oxidation of the sulfur present in the methyllanthionine ring, modification that seems to be catalyzed by a luciferase-like Flavin-dependent monooxygenase protein³⁹.

Class III and IV. Until now, only two types of PTM other than dehydration and cyclization were identified in class III and IV. They are disulfide bond formation and glycosylation. The sulfur bridge between cysteines has to be catalyzed by a disulfide isomerase, but the gene for the enzyme has yet to be identified, since it's not present in the cluster²⁹. The glycosylation is done to a tryptophan of labionin-containing NAI-112 and is the first and only reported case of a glycosylated residue in lanthipeptides, making this PTM exclusive of class III peptides⁴⁰.

Overall, besides the characteristic ring and dehydrated amino acids, the additional PTMs present in all classes of lanthipeptides contribute to a broader diversity of molecules and are sometimes essential for their characteristic bioactivities.

1.5 Immunity proteins

Many lanthipeptides are lantibiotics and have potent antibiotic activity that can also be harmful for the producing organism. Therefore, they need to be handled by special proteins also encoded within the biosynthetic clusters and that constitute their self-immunity system.

For class I. Lantibiotics like nisin act by interacting with lipid II. The strategy found by many organisms towards immunity is to limit said interaction with their cell wall. One mechanism relies

on a complex of three proteins for the efflux of the mature peptide present in the cytoplasm. Those proteins are encoded by *lanFEG*. NisF and NisE show strong homology to members of the family of ATP-binding cassette (ABC) transporters⁴¹. NisG encodes a hydrophobic protein that allocates this complex to the membrane. Although ingenious, the LanFEG transporter system is not a universal mechanism of immunity, as some organisms lack this locus in their biosynthetic clusters. Another strategy for autoimmunity relies on the sequestering of the lantibiotic in the extracellular space. This is achieved by a lipoprotein anchored to the cytoplasmic membrane. These proteins usually receive the designation LanI in the cluster¹⁸. Although they can act alone, these lipoproteins can also contribute cooperatively to full immunity with the respective LanFEG-related complexes¹⁸. Again, this lipoprotein is not present in all clusters, and the molecular means by which the producing strains achieve autoimmunity without any of these genes are still not clear⁷.

For class II. Several class II lanthipeptide clusters also encode LanEFG proteins and LanI. Although a special case occurs in this class of peptide. As in LanEFG and LanI cooperation towards full immunity, NukH, an immunity protein for Nukacin produced by *Staphylococcus warneri* ISK-1, works in cooperation with the NukFEG complex. This new immunity protein appears to be similar to other LanI except in some key aspects. except in some key aspects: i) NukH is rather small compared to the other LanI peptides, ii) NukH seems to be anchored to the cytoplasmic side of the membrane and iii) NukH appears to have broader substrate specificity⁴².

Another unique type of immunity in class II is found in the cinnamycin producing strain, *Streptomyces cinnamoneus* DSM 40646. This lantibiotic exerts its function by binding to phosphatidyl ethanolamine, instead of lipid II. The mechanism developed by this organism towards immunity involves methylation of the cinnamycin target by a methyltransferase. This enzyme is encoded by the *cinorf10* gene and its activity was confirmed with heterologous expression in *E. coli* which resulted in methylated phosphatidyl ethanolamine accumulation and resistance to the cinnamycin related lantibiotic, duramycin⁴³. Interestingly, this immunity mechanism does not appear to be universal as a homologue of *cinorf10* was not found in the genome of *Streptomyces cinnamoneus* ATCC 12686 the producer of duramycin. How the organism defends itself against the duramycin it produces is currently not known⁷.

1.6 Regulation

Lanthipeptide synthesis is an energy intensive process that needs to be tightly regulated. Several mechanisms were developed by producing organisms to control peptide biosynthesis. The prototypical lanthipeptide nisin is regulated by a two-component system NisRK¹³. NisK is a cytoplasmic membrane bound kinase that recognizes nisin. When the mature peptide binds to NisK, a cascade of events is unfolded. NisK starts with autophosphorylation, then following the phosphorylation of the second component of this system, NisR. NisR is a transcriptional factor that binds to the promoter regions of the biosynthetic cluster and induces synthesis¹³. The *nisRK* operon is independent of this nisin autoregulation, as NisK and NisR seem to be continuously produced⁴⁴. Several other lanthipeptides like subtilin, mersacidin and salivaricin A were reported

to have a similar auto regulated two-component system⁴. Another adopted strategy by producing organisms relies on a single regulator protein. In the case of lactacin 3147, a constitutively expressed transcription factor controls biosynthesis and a second transcriptional repressor encoded by *ltnR* regulates expression of the immunity proteins⁴⁵. A combination of the two was also confirmed. In the case of mersacidin it was reported that peptide biosynthesis is regulated solely by MrsR1 and immunity proteins were regulated by a two component system, MrsR2/K2⁴⁶. Allied to this mechanisms, there are also several other factors to consider. Cell cycle seems to play a role in biosynthesis regulation. Several regulatory systems involved in different cell phases may also act on the biosynthetic clusters. Growing conditions may also dictate the expression of peptides. For instance production of the virulence factor cytolysin is stimulated in the presence of the target cells⁴. Regulation can be therefore highly intricate and the discovery of new mechanisms is to be expected.

1.7 Genome Mining

Next generation sequencing and turn of the century genome sequencing efforts allowed for the exponentially increasing publicly available number of bacterial genomes. Meaning that in addition to traditional microbiological approaches, *in silico* screening strategies may now be employed in the discovery of new natural products. Taking advantage of the highly conserved nature of lanthipeptide biosynthetic enzymes, their organization in clusters and direct connection between peptide and corresponding gene (Figure 7), genome mining has proved highly prolific in identifying and isolating new compounds directly from producing organisms or through heterologous production systems – a bottom-up approach. The new information gathered also facilitates new discoveries in a more traditional approach. A top-down discovery, when a relevant biological activity is firstly identified, then a screening or even sequencing of the producer organism genome allows to identify the possible biosynthetic machinery at play. It's a mixture of these methods that allowed for the explosion of new and varied RiPPs⁴⁷.

There are some limitations for genome mining. Firstly, there's very limited information about the biological activity of the natural product newly identified *in silico*. Secondly, genomic screening is limited to the knowledge available about protein families and activities. If new biosynthetic enzymes with completely unrelated structures to the ones already identified are at play, they will remain unknown. Only a continued effort to identify new compounds will reduce the present gap in knowledge. And finally, the precursor peptide's gene and some of its biosynthetic machinery might not be neatly organized in a characteristic cluster making it difficult to associate a precursor peptide to its biosynthetic machinery.

Nevertheless, these characteristic clusters of RiPPs and lanthipeptides biosynthesis makes them excellent targets for bioengineering explorations. Recent efforts using PSI-Blast to identify new homologue biosynthetic proteins in all publicly available genomes aided with readily available genome mining tools developed specially for RiPPs, among them being BAGEL2⁴⁸ and antiSMASH⁴⁹ that screen for putative peptides based on a large set of search criteria, were returned with positive results.

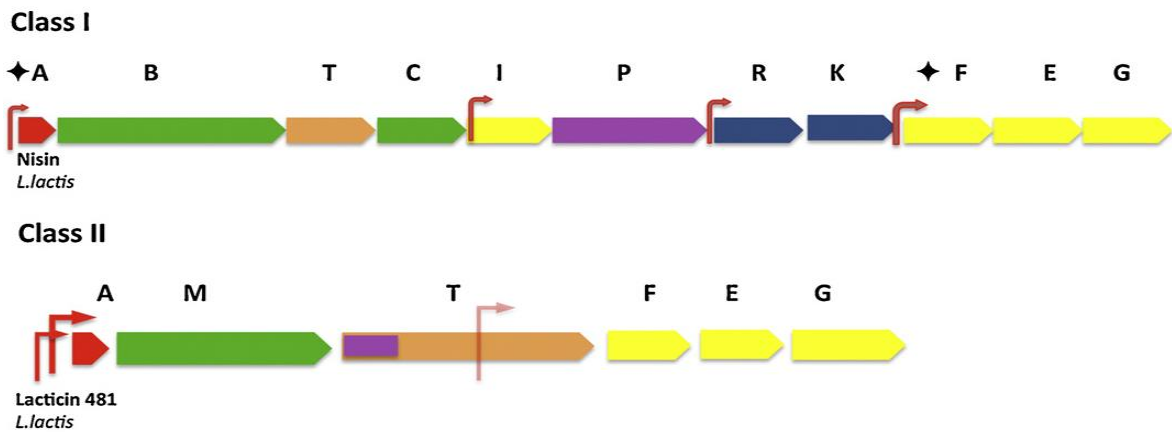


Figure 7 Gene cluster representation from class I and class II lanthipeptides. Example from Nisin and Lactacin 481. In green are the genes responsible for ring formation and lanthipeptide classification. In class I those are the lanB and lanC genes. In class II is the lanM gene. lanT and lanP genes are usually responsible for export and cleavage of the precursor peptide in class I. A bifunctional lanT assumes that role in class II. In red are the precursor peptides. In yellow immunity proteins. lanRK is the two component regulator of nisin. Image adapted from Alkhatib et al, 2012⁵⁰.

1.8 Lanthipeptides heterologous expression in *Escherichia coli*

Escherichia coli is a well-established model for heterologous expression and the organism of choice in a wide range of applications. There are therefore many molecular tools and readily available protocols for highly efficient protein expression. The principal for heterologous expression is very straightforward. After selection of the gene of interest follows cloning in an appropriate vector. After induction, the protein is purified in the varied ways developed to so far, one of them being IMAC (Immobilized metal affinity chromatography). After this, characterization ensues. However, many things can go wrong throughout the process. Poor growth of the host organism, protein inactivity, inclusion bodies formation or lack of protein production are the most common problems encountered⁵¹. Giving the wide distribution of the biosynthetic clusters in all domains of life, some of these setbacks are to be expected if heterologous expression in this common model is attempted.

The first reported lanthipeptide modified in *E. coli* dates back to 2005, with the works of Nagao et al where they coexpressed a His-tagged nukA peptide with the class II nukM enzyme in a single plasmid vector⁵². Class I peptides have also been modified within a *E. coli* model, where nisA was cloned together with nisB in a vector and nisC on a second plasmid, yielding a fully modified nisin precursor peptide²⁰. The *E. coli* expression system is now considered an appropriate method to produce lanthipeptides and in conjunction with in vitro studies contributed to furthering the knowledge of these biosynthetic clusters.

1.9 The third domain of life – Archaea

Classical systematics relying not only on complex morphological but also physiological studies and fossil records give a limited view on the evolution of life on earth. What was at first an incomplete view of life that classified all organisms either as a plant or as an animal, with the discovery of

bacterial organisms a new equally incomplete view of life emerged as eukaryotic vs prokaryotic. The availability of molecular tools by mid twentieth century made it possible to establish a real evolutionary relationship between organisms and give a more appropriate order to the diversity of life. Carl R. Woese believed this by stating that an 'organism's genome seems to be the ultimate record of its evolutionary history'⁵³ and shook the taxonomic world with a groundbreaking work published in 1977. Using ribosomal RNA sequence as a tool to detect phylogenetic relations between organisms revealed the existence of a distinct third group, thought until then to be part of the prokaryotes based on classical methodologies. They termed that group Archaea. Later on and based on these molecular studies of evolutionary relationships, Woese et al (1990) proposed the revamping of the outdated and misleading views on Taxa with a new formal system of classification based on the three domains of life: Eukarya, Archaea and Bacteria⁵⁴. Although being quickly attributed to extreme environments, the Archaea are now known to be metabolically diverse organisms coexisting in various environments with Eukarya and Bacteria, contributing even with considerable biomass in some cases. Genome sequencing revealed that Archaea contain a mix of bacterial and eukaryotic features. Their information processing mechanisms resemble the eukaryotic ones, while the core metabolic functions are similar to bacteria. This is not a strict rule, as many exceptions exist. Similar information processing mechanisms means, for instance, that the archaeal RNA polymerase is closely related to RNA polymerase II from the eukaryotic counterpart. Not only are the enzymes similar, they also require basal factors for promoter recognition, possessing even homologues for the eukaryotic TATA-box binding protein and transcription factor B known from the eukaryotic systems. Despite that, the transcription machinery appears to be much simpler in Archaea. There seems to be also the presence of bacterial like transcription regulators in archaeal genomes meaning that regulation is not entirely eukaryotic like⁵⁵. Another area of interest in Archaea is DNA recombination and repair. The potential for archaeal studies in the subject was made evident by the discovery of a new family of type II topoisomerases in Archaea with homology to proteins involved in meiotic recombination⁵⁶. The study of these enzymes opens the possibility of gaining a better understanding of the intricate process of meiosis by eukaryotes. Given their characteristics, various other interesting applications are at play. Archaea have also been used to study DNA replication⁵⁵. Structural characterization of archaeal proteins that evolved for high temperatures and saline environments allow for data gathering useful in the study of a range of possible protein interactions in various organisms. And finally, there is the biotechnological application where the extremophile characteristics of these organisms may allow for a higher yield of a given product. But these studies are only possible with the development of the correct molecular and biochemical tools and the existence of a proper model organism. Halophilic archaea are able to grow in hypersaline environments. The coping mechanism for these extreme conditions rely on accumulating high cytoplasmic concentrations of potassium. For the enzymatic machinery to function in this high salt concentration within the cell, evolution worked towards reducing hydrophobicity of these proteins and increasing the number of acidic residues on the protein surface⁵⁷. The negative charge rich surface interacts with cations in solution, allowing for a solvated state of the protein. This ability can be a set back with the use of model organisms like *E. coli* and their extensively developed manipulation tools. Halophilic enzymes in *E. coli* environment can be misfold and aggregate in inclusion bodies that would then require, besides the purification steps, the denaturation followed by protein refolding in appropriate conditions⁵⁸. This can now be circumvented with the use of the model organism for Archaeal genetics, *Haloferax volcanii*.

The development of selectable markers⁵⁹, gene knockout systems⁶⁰ and promoter characterization allowed for the development of specific strains and plasmid vectors essential for protein studies⁶¹. *Haloferax volcanii* revealed itself to be even a greater choice for the production of halophilic proteins for the various biotechnological purposes, with the development of suitable expression protocols⁶².

1.10 *Haloferax mediterranei*

Haloferax mediterranei is a halophilic organism of the *Halobacteriaceae* family, domain Archaea. It was first isolated in 1980 from salt water evaporation ponds near Alicante, Spain⁶³. Being a halophilic organism is capable of growth in a varied range of high salt concentrations. The formation of opaque to pink colonies is associated with pigment production that in the first isolation report stated to vary according to salt concentration. A seemingly unique characteristic of this particular organism⁶³. Other interesting attributes noted were the absence of complex nutrient requirements, rapid growth rates and nutritional versatility all relevant factors for growth in the microbiologist laboratory⁶³.

Studies on a varied number of metabolites from this archaeon and a potential for industrial application led to the necessity of its genome sequencing in 2012⁶⁴. The genome of *Haloferax mediterranei* ATCC 33500 consists of one chromosome and three megaplasms. Comparison with the archaeal model genome, *Haloferax volcanii*, reveals high homology between the chromosomes and a divergence on the plasmids⁶⁴.

1.11 Work Objectives

Lanthipeptides are an ever growing family of natural products. Lanthipeptide's activity was once thought to be restricted to antibiotics, but since then the activity spectrum and clinical applications of these peptides have been extended. *In silico* analysis identifies their biosynthetic enzymes in all domains of life. However, and so far, no lanthipeptides have been described in the Domain Archaea. Focused in class II enzymes genes present in Archaea, this work has focused on the identification of new biotechnologically interesting molecules and enzymes in this underexplored domain.

To that end, three primary objectives were established:

- i) Identification of lanthipeptides clusters in Archaea, using bioinformatics tools and databases publicly available.
- ii) Characterization of the identified clusters followed by comparison with other known lanthipeptides clusters.
- iii) Identification of the dehydration pattern in putative lanthipeptides from *H. mediterranei* ATCC 33500, through heterologous expression, peptide purification and mass spectrometry analysis.

2. Materials and Methods

2.1 Bioinformatics

Novel putative lantibiotic modifying enzymes were identified using Position-Specific Iterated BLAST (PSI-BLAST) of the lantipeptide synthetase CylM with accession number AAK67266.1 in a search at National Center for Biotechnology Information (NCBI) Database (<https://www.ncbi.nlm.nih.gov/>). The initial Blast search was limited to the Archaea domain and the standard algorithm parameters were used. Two more iterations followed. All relevant hits were examined.

Manual search for putative precursor peptides was done using ORFfinder (<https://www.ncbi.nlm.nih.gov/orffinder/>).

The genomes encoding for the identified LanMs were analyzed using Bagel3⁴⁸ (<http://bagel.molgenrug.nl/index.php/main>) for the purpose of identifying the respective gene clusters. TMHMM (<http://www.cbs.dtu.dk/services/TMHMM>) was used to predict transmembrane helices and InterPro⁶⁵ for protein domain analysis .

The sequence Similarity Network (SSN) was obtained by the online Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST)⁶⁶ using as reference the Pfam of LanM enzymes (PF13575), an e value of -10 and an alignment score corresponding to 35% identity. The SSN was then visualized using Cytoscape⁶⁷.

2.2 Media and growth conditions

Haloferax mediterranei ATCC 33500 was grown at 37°C in a medium containing 156 g L⁻¹ NaCl, 13 g L⁻¹ MgCl₂·6H₂O, 20 g L⁻¹ MgSO₄ · 7H₂O, 1 g L⁻¹ CaCl₂ · 6H₂O, 4 g L⁻¹ KCl, 0.2 g L⁻¹ NaHCO₃, 0.5 g L⁻¹ NaBr, 5 g L⁻¹ Yeast extract and 1 g L⁻¹ Glucose.

Haloferax volcanii H1424 was grown in Hv-YPC medium (144 g L⁻¹ NaCl, 21 g L⁻¹ MgSO₄·7H₂O, 18 g L⁻¹ MgCl₂·6H₂O, 4.2 g L⁻¹ KCl, 5 g L⁻¹ yeast extract, 1 g L⁻¹ peptone from meat, 1 g L⁻¹ casamino acids, 12mM Tris, pH 7.5, 3mM CaCl₂ added after sterilization) at 45°C. For solid media, agar was added to a concentration of 15 g L⁻¹ and was heat dissolved before adding the peptone from meat, yeast extract and the casamino acids. The medium was then autoclaved and CaCl₂ added. When required, thymidine was added to a concentration of 40 µg mL⁻¹.

Escherichia coli DH5 α and BL21-Gold (DE3) strains were grown in Luria Bertani medium containing the appropriate selective markers.

2.3 DNA extraction

The extraction of DNA from *Haloferax mediterranei* ATCC 33500 was performed with the DNeasy Blood and Tissue kit (Quiagen) using the protocols recommended for Gram negative and Gram positive bacteria. After electrophoresis analysis, both protocols proved to be effective.

Plasmid DNA was extracted with the GeneJET Plasmid Miniprep Kit (Thermo), according to manufacturer's instructions.

2.4 *E.coli* co-expression system

2.4.1 *medA* and *medM* amplification and plasmid construction for *E. coli* expression

The pRSFDuet-1 plasmid allows co-expression of two genes due to its two multiple cloning sites (MCS). The *medA* genes were amplified from *H. mediterranei* genomic DNA with primers with the appropriate adaptors to allow the expression of a N-terminal His6-tagged precursor peptide at the MCS-1. *medM* genes were amplified for MCS-2 insertion. The primers used are listed on Table 1. Reaction components and conditions are described in table 2 and 3, respectively.

Table 1 Primers used for gene amplification by PCR for *E. coli* expression. The enzymes shown were the enzymes used for cloning and the Ta represents the temperature of annealing used in each PCR reaction.

Gene	Enzymes	Primers	Ta	Amplification (bp)
<i>medA1</i>	<i>Bam</i> HI <i>Not</i> I	5' atgaggatccgctcggtagtagtactcgacatcg 3' 5' cgatgcggccgcttagaaatcgaggtatggca 3'	66°C	170
<i>medM1</i>	<i>Nde</i> I <i>Xho</i> I	5' agtacatatgacacagcagcttgacgc 3' 5' cgatctcgagttactccagcagaagcacacag 3'	63°C	3161
<i>medA2</i>	<i>Bam</i> HI <i>Not</i> I	5'atgaggatccgtccctccgtgtcccta 3' 5'cgatgcggccgcttagcatatcagacatctgaa 3'	66°C	227
<i>medM2</i>	<i>Nde</i> I <i>Xho</i> I	5' agtacatatgaaccgctgtacacgatg 3' 5' cgatctcgagttactcaagcaagagaacgga 3'	62°C	3209
<i>medA3</i>	<i>Bam</i> HI <i>Not</i> I	5' atgaggatccgtcagcatatcagacatctg 3' 5' cgatgcggccgcttactgagtagtccgtgatgca 3'	65°C	212
<i>medM3</i>	<i>Nde</i> I <i>Xho</i> I	5' agtacatatggcggcggtatttactgag 3' 5' cgatctcgagttattccagcgtaacacgtt 3'	62°C	3239

Table 2 PCR reaction used for amplification of genes.

Reagents	Final concentration	Volume (μ L)
10x Reaction Buffer	1x	5
10 mM dNTPs mix	0.2mM	1
10 μ M Primer Fw	0.3 μ M	1.5
10 μ M Primer Rv	0.3 μ M	1.5
Template DNA	10 ng	1.0
NZYProof DNA polymerase 2.5 U/ μ L	0.05 U/ μ L	0.5
dH ₂ O	-	Up to 50

Table 3 PCR conditions used for gene amplification. * represents the Ta used in each reaction and that is shown in Table 1.

	Temperature	Time	
Initial denaturation	95°C	3 min	
Denaturation	95°C	30 s	30 cycles
Annealing	*°C	30 s	
Extension	72°C	60s/kb	
Final Extension	72°C	10 min	

After amplification, 1 µg of each PCR product and 1 µg of plasmid were digested with FastDigest enzymes (Thermo) according to the recommended protocol. After digestion, the fragments were purified from the agarose gel with the QIAquick Gel Extraction Kit (QIAGEN). The resulting digested PCR products and plasmids were ligated in a 3:1 molar ratio proportion with 1U of T4 DNA ligase and 50 ng of the vector (Thermo Fisher Scientific). The reaction was performed at 22°C for 1h; 5µl of the ligation used to transform chemically competent *E. coli* DH5α by heat shock. Clones were selected on LB agar plates containing kanamycin (100 mg mL⁻¹). Positive clones were screened by colony-PCR using the universal primers described for MCS1 or MCS2. All the positive plasmids were sequenced at STAB VIDA, Lda services, to confirm the absence of mutations.

2.4.2 Co-expression and purification of the *medA* and *medM* products

After transformation of *E. coli* BL21-Gold(DE3) with the plasmids constructed, single colony transformants were grown overnight at 37° in 5 mL of Terrific Broth (TB) medium supplemented with kanamycin (100 mg mL⁻¹). The cells were then transferred to 60 mL of fresh TB medium and grown at 37°C until O.D₆₀₀ ~0.7-0.8. At this stage, IPTG was added to a final concentration of 0.3mM. The culture was then incubated at 37°C and/or 18°C overnight. Cells were harvested by centrifugation at 4000 rpm for 15 minutes. The peptides were purified from *E. coli* pellets using a well-established protocol for His₆ tagged LanA peptides⁶⁸ and a 1 mL Hitrap chelating HP nickel affinity column (GE Healthcare). C4 Solid phase extraction (SPE) was used for sample desalting before mass spectrometry analysis, as described in section 2.6.

2.5 *Haloferax volcanii* expression system

2.5.1 *medA* and *medM* amplification and plasmid construction for *Haloferax volcanii* co-expression

The plasmid used, pTA1392, and the *Haloferax volcanii* H1424 strain were kindly provided by Professor Thorsten Allers of the Institute of Genetics, School of Biology from the University of Nottingham. The plasmid has only one promoter, followed by a MCS. In order to mimic the Duet system, the final plasmid was constructed in two steps: i) firstly, all *medA* and *medM* genes were amplified and cloned into pTA1392 MCS and ii) secondly, the pTA1392 MCS containing *medAs* was amplified and cloned in pTA1392 plasmids already containing *medM* genes. The DNA sequences of the genes of interest were amplified harboring the appropriate flanking sequences for multi cloning site insertion. The *medA* genes were individually inserted in a first plasmid using restriction enzymes allowing for the His₆Tag residues to be transcribed at the N-terminal of the peptides. The *medM* genes were inserted in the same plasmid using the appropriate restriction enzymes. In the second step, the promoter and the cloned *medA* genes were amplified by PCR with the appropriate flanking sequences using the newly constructed plasmids (pTA1392 with *medA* genes) as template. This entire sequence was then cloned in the plasmid containing the *medM* of interest. The amplification of genes, digestion and ligation was performed as described in section 2.4. All primers used are listed in Table 2. Positive colonies, ampicillin resistant cells (100 mg mL⁻¹), were screened by colony-PCR. The absence of mutations in all the genes inserted in the plasmids was confirmed by sequencing at STAB VIDA, Lda.

Table 4 Primers used for gene amplification by PCR for *Haloferax volcanii* expression.

Gene	Enzymes	Primers	Ta	Amplification (bp)
<i>medA1</i>	<i>NspI</i> <i>BamHI</i>	5' atgaacatgtggtcggtagtactcgacatcg 3' 5' cgatggatccttagaaatcgaggtatggca 3'	62°C	169
<i>medM1</i>	<i>NdeI</i> <i>NheI</i>	5' agtacaatgatgacacagcagcttgagc 3' 5' cgatgctagctcactccagcagaagcacacag 3'	64°C	3164
<i>medA2</i>	<i>NspI</i> <i>NotI</i>	5' atgaacatgtggtccctccgtgtcccta 3' 5' cgatcgccgctcagcatatcagacatct 3'	66°C	228
<i>medM2</i>	<i>NdeI</i> <i>BamHI</i>	5' agtacaatgaaccgctctacacgatg 3' 5' cgatggatccctactcaagcaagagaacgga 3'	62°C	3209
<i>medA3</i>	<i>NspI</i> <i>BamHI</i>	5' atgaacatgtggtcagcatatcagacatctg 3' 5' cgatggatccctactgagtatccgtgatgca 3'	62°C	211
<i>medM3</i>	<i>NdeI</i> <i>NheI</i>	5' agtacaatggcggcgctatttactgag 3' 5' cgatgctagctattccagcgtaacacgtt 3'	62°C	3239
Ptna promoter + <i>medA1</i>	<i>NheI</i> <i>BamHI</i>	5' cgatgctagccgctctcgaagctgttc 3' 5' cgatggatccttagaaatcgaggtatggca 3'	64°C	404
Ptna promoter + <i>medA2</i>	<i>BamHI</i> <i>NotI</i>	5' cgatggatcccgctctcgaagctgttc 3' 5' cgatcgccgctcagcatatcagacatct 3'	65°C	463
Ptna promoter + <i>medA3</i>	<i>NheI</i> <i>BamHI</i>	5' cgatgctagccgctctcgaagctgttc 3' 5' cgatggatccctactgagtatccgtgatgca 3'	64°C	446

2.5.2 *Haloferax volcanii* H1424 transformation

The plasmids constructed in the previous section were used to transform *H. volcanii* H1424. Transformation was performed by the spheroplast formation in polyethylene glycol 600 method described by Cline and Doolittle⁶⁹. Selection is possible due to the highly engineered nature of the strain H1424. This strain lacks the *pyrE2* gene encoding for an enzyme essential for pyrimidine *de novo* biosynthesis. Successful transformation reintroduces this gene in the organism, allowing growth in medium lacking thymidine. All strains obtained from transformation are summarized in table 7.

Table 5 Summary of all strains obtained from transformation.

Strain	Description
H1424_medA1	pTA1392 containing <i>medA1</i> gene N-terminally fused to an His ₆ Tag
H1424_medM1	pTA1392 containing <i>medM1</i> gene
H1424_medA2	pTA1392 containing <i>medA2</i> gene N-terminally fused to an His ₆ Tag
H1424_medM2	pTA1392 containing <i>medM2</i> gene
H1424_medA3	pTA1392 containing <i>medA3</i> gene N-terminally fused to an His ₆ Tag
H1424_medM3	pTA1392 containing <i>medM3</i> gene
H1424_medM1A1	pTA1392 containing <i>medM1</i> gene and <i>medA1</i> gene N-terminally fused to an His ₆ Tag
H1424_medM2A2	pTA1392 containing <i>medM2</i> gene and <i>medA2</i> gene N-terminally fused to an His ₆ Tag
H1424_medM3A3	pTA1392 containing <i>medM3</i> gene and <i>medA3</i> gene N-terminally fused to an His ₆ Tag

2.5.3 MedAs expression in *Haloferax volcanii* H1424

The protocol described by Strillinger et al⁶² was adopted for our convenience. *Haloferax volcanii* transformed with the desired plasmid was initially grown in 20mL of Hv-YPC broth at 45°C and 180 rpm for 24 hours. This initial culture was then diluted to a final OD_{610nm} of 0.005 in fresh medium, for a final volume of 100mL. The cultures were left to grow in the same conditions until OD_{610nm} of 0.3. When OD_{610nm}=0.3, tryptophan was added to a final concentration of 6mM. After 16 hours of incubation, tryptophan was added again to a final concentration of 9mM and the cultures were grown for another 4 hours. At this time, the cells were collected by centrifugation at 4000 rpm for 15 minutes and suspended in 7 mL of Buffer A (20 mM HEPES pH7.5, 2M NaCl, 20 mM imidazole). The cells were then lysed by sonication on ice until the suspension was no longer as turbid, and centrifuged at 10000 rpm for 15 minutes. The supernatant obtained was filtered by a 0.45 µm PES filter. The His₆Tag-LanA peptides were purified with a 1 mL Hitrap chelating HP nickel affinity column (GE Healthcare). This column was first equilibrated with 5 column volumes (CVs) of buffer A. The filtered supernatant was loaded to the equilibrated column and the column

was washed with 5 CVs of buffer A. The peptides were then eluted with 2 mL of Buffer A containing 100mM imidazole.

2.6 MALDI-TOF Mass spectrometry analysis

After IMAC purification of peptides produced in *E. coli* and *H. volcanii*, the samples were desalted using ZipTip C₁₈ or by solid phase extraction (SPE) with a C4 column, before MALDI-TOF analysis in a 4800 Plus MALDI TOF/TOF Analyzer (AB SCIEX).

When ZipTip was used, the ZipTip was equilibrated with 100% acetonitrile (ACN) followed by 0.1% trifluoroacetic acid in mqH₂O (0,1% TFA). After peptide binding to the ZipTip, it was washed with 0.1% TFA. The elution of the peptides was performed with a saturated solution of α -cyano-4-hydroxycinnamic acid (10mg mL⁻¹; CHCA) prepared in 50%ACN acidified with 0.1 % TFA directly to the MS plate.

Desalting by SPE involved the equilibration of the C4 column with 100% ACN followed by wash with 0.1 % TFA. Then, the sample was loaded and washed twice with 0.1% TFA. Elution was performed with 70% ACN. After evaporation, the sample was suspended in 0.1% TFA and mixed with the saturated solution of CHCA for MS analysis.

3. Results and discussion

3.1 Identification of CylM homologues in Archaeal genomes

The enzyme CylM was used as the query sequence for mining of homologues in the Archaea domain. CylM is a member of the type 2 lantibiotic biosynthesis protein LanM family. As such, CylM is a bifunctional enzyme with a dehydratase domain in its N-terminal region, responsible for the dehydration of Serine and Threonine residues, and a cyclase domain in the C-terminal region responsible for the formation of the lanthionine/methyllanthionine rings. This cyclase domain belongs to the lanthionine synthetase C-like protein family.

Position-Specific Iterative (PSI)-BLAST is a protein database search program that builds off the initial alignment, that is similar to a BLASTp one⁷⁰. After the first iteration a statistically significant alignment is produced from the highest scoring hits of the initial run and this new profile, or position-specific score matrix (PSSM) where highly conserved residues receive the highest scores. It is used instead of the original substitution matrix for another sequence search. Each iteration translates into a refinement of the PSSM. This process allows the discovery of related sequences that a simple BLASTp run wouldn't be able to do.

The PSI-Blast results gave a total of 25 hits for possible LanM enzymes in Archaeal genomes. From these, four hits were discarded based only on the sequence length. Smaller proteins were probably homologues of the LanC like domain of our query sequence. These proteins are LanCs and not LanMs that are capable both of dehydration and cyclization. Thus, the proteins belonging to *Haloterrigena mahii*, *Halophilic archaeon* J07HB67 and both of *Halopiger djelfmassiliensis* were not considered for further analysis.

All the significant hits of putative LanM proteins in the Archaea domain were exclusively detected in organisms from the class Halobacteria and are represented in Table 3. The most distinctive feature of these organisms is their requirement for high salt concentrations. Some species are able to grow at concentrations as low as 1 M NaCl, but most of them grow best between 3.5 and 4.5 molar concentrations⁷¹. The mechanism behind the tolerance to this environment seems to be the accumulation of KCl in the cell and therefore their enzymatic machinery evolved to have the highest activity at high levels of KCl⁷¹. Nearly all species possess distinctive red, orange or pink pigmentation.

Table 6 CylM homologues identified by PSI-Blast. All organisms belong to phylum Euryarchaeota, class Halobacteria.

			Organism	Protein accession number	Genome RefSeq	Locus tag
Halobacteriales	Halobacteriaceae	Haladaptatus	Haladaptatus cibarius D43	WP_049969213.1	GCF_000710615	HL45_RS00765
				WP_049972812.1		HL45_RS19115
				WP_049971341.1		HL45_RS12075
			Haladaptatus paucihalophilus DX253	WP_007983023.1	NZ_AQXI000000000	B208_RS0118315
	Halorculaceae	Haloarcula	Haloarcula argentinensis DSM 12282	WP_049943790.1	NZ_AOLX000000000	C443_RS00600
		Halomicrobium	Halomicrobium mukohataei DSM 12286	WP_015761762.1	GCF_000023965	HMUK_RS03740
	unclassified Halobacteriales		halophilic archaeon J07HB67	ERH12559.1 424 aa	GCA_000416105	J07HB67_01580
halophilic archaeon J07HX5			ERG89078.1	GCA_000415945	J07HX5_01230	
Haloferacales	Haloferacaceae	Haloferax	Haloferax mediterranei ATCC 33500	WP_004056317.1	GCF_000306765(Plasmid pHM300)	HFX_RS15990
				WP_004056313.1	GCF_000306765(Plasmid pHM300)	HFX_RS16005
				WP_004057361.1	GCF_000306765 (Chromosome)	HFX_RS04210
			Haloferax alexandrinus JCM 10717	WP_049968936.1	GCF_000723845	BN984_RS18090
			Haloferax prahovense DSM 18310	WP_008093839.1	NZ_AOLG000000000	C457_RS08840
			Haloferax denitrificans ATCC 35960	WP_004967897.1	NZ_AOLP000000000	C438_RS03580
			Haloferax gibbonsii ATCC 33959	WP_050460258.1	NZ_CP011949 (Plasmid pHG2)	ABY42_RS17125
			Haloferax sp. ATB1	WP_042665535.1	GCF_000730175	ATB1_RS15910
			Haloferax larsenii JCM 13917	WP_007538897.1	GCF_900109695	C455_RS00665

Table 8 CylM homologues identified by PSI-Blast. All organisms belong to phylum Euryarchaeota, class Halobacteria (continuation of Table 8).

			Organism	Protein accession number	Genome RefSeq	Locus tag
Natrialbales	Natrialbaceae	<i>Natrinema</i>	<i>Natrinema</i> sp. J7-2	WP_014862558.1	NC_018224	NJ7G_RS00200
			<i>Natrinema gari</i> JCM 14663	WP_049910399.1	NZ_AOIJ000000000	C486_RS10070
		<i>Halobiforma</i>	<i>Halobiforma lacisalsi</i> AJ5	WP_007143049.1	NZ_CP019286 (plasmid pHLAJ5I)	CHINAEXTREME_RS20780
		<i>Natronorubrum</i>	<i>Natronorubrum tibetense</i> GA33	WP_006092971.1	GCF_000383975	NATTI_RS0124325
		<i>Halopiger</i>	<i>Halopiger djelfimassiliensis</i>	WP_049921103.1 654 aa	NZ_CBMA000000000	TX80_RS01480
				WP_049921102.1 402 aa	NZ_CBMA000000000	TX80_RS01475
		<i>Haloterrigena</i>	<i>Haloterrigena mahii</i>	WP_082917683.1 419 aa	GCF_000690595	HTG_RS00500
		unclassified Natrialbaceae	<i>Natrialbaceae archaeon</i> JW/NM-HA 15	WP_086889213.1	GCF_002156705	B1756_RS14655

3.2 Conserved residues between CylM and the putative LanM

Several residues are conserved not only across LanM but also between LanM and LanC given that LanM enzymes possess a LanC like domain responsible for the cyclization reaction. The importance of several of these amino acids was confirmed through mutational studies and the X-ray structure of CylM^{7,25}. A number of conserved residues is expected given the nature of the PSI-BLAST search. However, the presence of the specific conserved residues known to be involved in the dehydration and cyclization reactions was confirmed for all the enzymes through Multiple Sequence Alignment (MSA) (Figures 8 and 9).

Residues involved in Dehydration

In CylM, the amino acids that are thought to be responsible for the activation of the phosphate to be transferred from ATP to the substrate in the phosphorylation step are Asp252 and His254. However, Lys274, Asp347, His349, Asn352, Asp364 and Glu366 are also involved in phosphorylation⁷. Glu366 is thought to stabilize ATP during the reaction and Asn352 with Asp364 might interact with ATP through a magnesium cation. In some reported cases, Gln replaces His349. All of these residues were present in the LanM of Archaea. Moreover, it was found that the majority of the proteins contain the Gln349 instead of His (Figure 8). Other residues were reported to be important for LanM activity based on mutagenesis studies. On CylM, these residues are Tyr330, Glu555 and Gln611. In Archaea, Tyr330 is conserved across all enzymes, but Glu555 is not. Although there is no conservation of the glutamic acid, other charged residues like arginine, lysine and aspartic acid are present (Figure 8). Gln611 is conserved in all enzymes but one encoded in *Haloferax larsenii* JCM 13917 genome, where instead of a glutamine there is a glutamic acid. Arg506 and Thr512 are important residues for phosphate elimination, the second step of the dehydration reaction and they are both conserved in all proteins.

Residues involved in Cyclization

The cyclase domain in CylM shares with NisC the α,α -barrel fold that encapsulates a Zn^{2+} cofactor by interaction with one histidine and two cysteine residues²⁵. Mutational studies revealed their importance for cation binding and consequently cyclization. These residues in CylM are located in Cys875, Cys911 and His912. In many LanMs (first reported for ProcM) instead of this motif, they contain three cysteine ligands for zinc. Euryarchaeota proteins were found to cluster almost entirely within this three cysteine motif group (Figure 9), which is present in all enzymes, except for *Haladaptus cibarius* D43 (WP_049972812.1). In the latter, no cyclization domain was detected.

LanC cyclases and the LaM cyclase domains appear to have other conserved His residue⁷: His212 for NisC and His790 for CylM. This residue is not conserved in the Archaeal LanM.

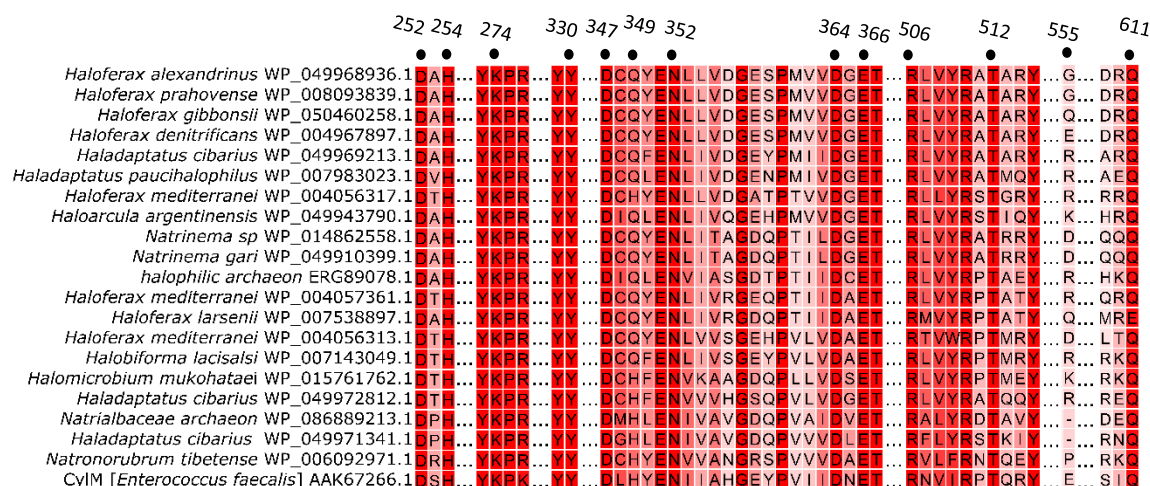


Figure 8 Multiple sequence alignment of the identified LanMs. Analysis showed strong conservation of the residues involved in dehydration.

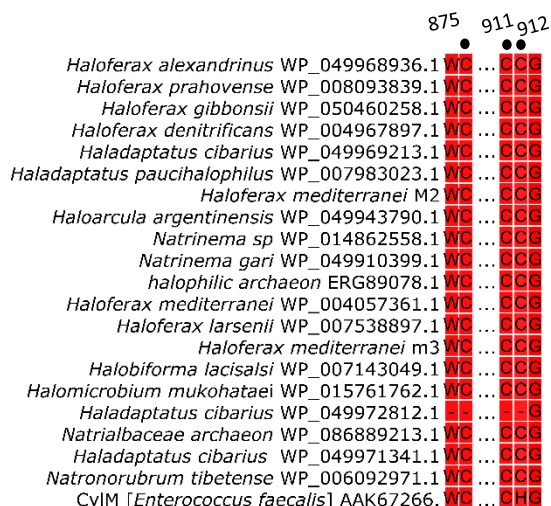


Figure 9 Multiple sequence alignment of the identified LanMs shows conservation of the three cysteine ligands for zinc.

The conservation of the three cysteine residues, the lack of conservation for the His790 residue and the adaption of these enzymes to high salt concentrations might be the reasons why they all seem to cluster together when a sequence similarity network analysis was performed (Figure 10). A SSN allows the visualization of sequence relationships between individual members of a large protein family and helps map out members of a family that are not well characterized. In Figure 10, we can clearly outline a cluster composed by the archaeal LanM enzymes.

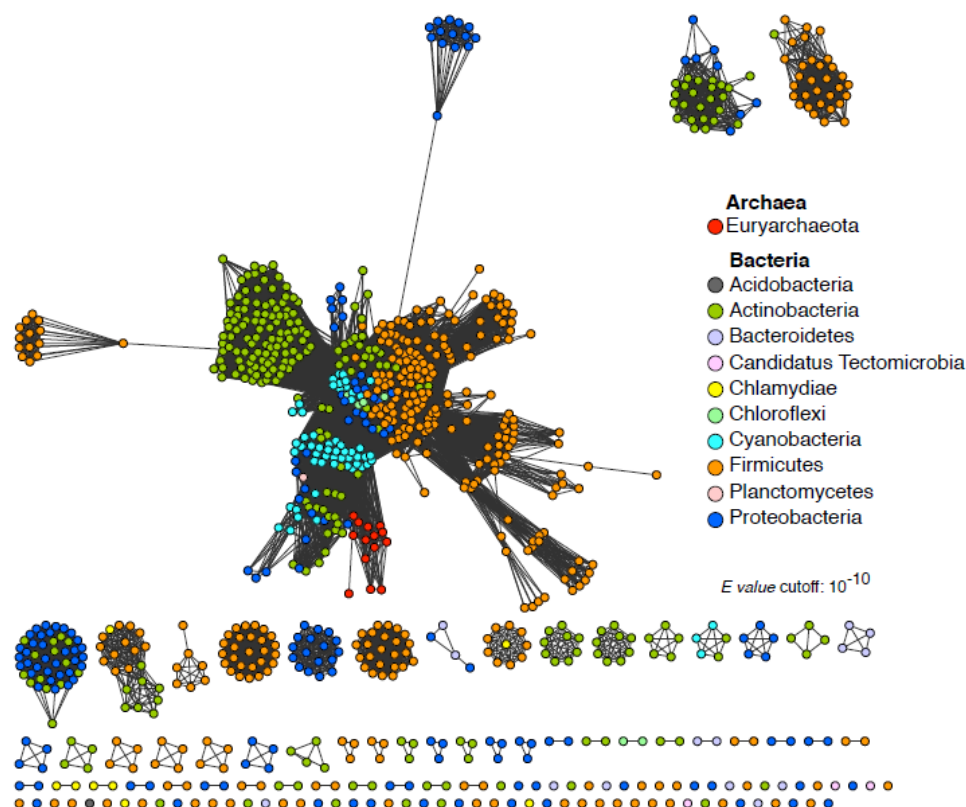


Figure 10 Sequence similarity network of LanMs generated using EFI-EST and visualized in Cytoscape with an alignment score threshold of 110 (~35% sequence identity).

3.3 Search and analysis of the putative archaeal precursor peptides

Genome mining tools rely on a set of parameters for the identification of the putative precursor peptides based on the knowledge involving the peptide sequence and the known biosynthetic clusters organization. These genome-mining tools are of utmost importance. However, very often they do not detect the genes encoding precursor peptides. This is mainly due to the fact that they are normally small genes that are not annotated in the publicly available genomes. Nevertheless, new promising tools are being developed towards mitigating these limitations⁷² that might prove highly effective in future research.

The majority of the precursor peptides are encoded near their corresponding modification enzyme. Thus, when a putative peptide was not annotated nearby the protein detected, we searched for their presence “manually” using ORFfinder. This allowed us to identify open reading frames that coded for small peptides containing serine, threonine and cysteine residues that are essential for lanthionine ring formation. Assuming that this expression is highly regulated and that LanA and LanM can be coexpressed in the same operon⁷, peptides encoded in the same

orientation of LanM were favored. With this approach, it was possible to detect *lanA* genes in all the archaeal genomes with *lanM* genes. We then compared all of the peptides in order to evaluate if they share some conserved motifs. This analysis could also help in the detection of a pattern that could allow the differentiation between their core and leader sequences (Figure 11).

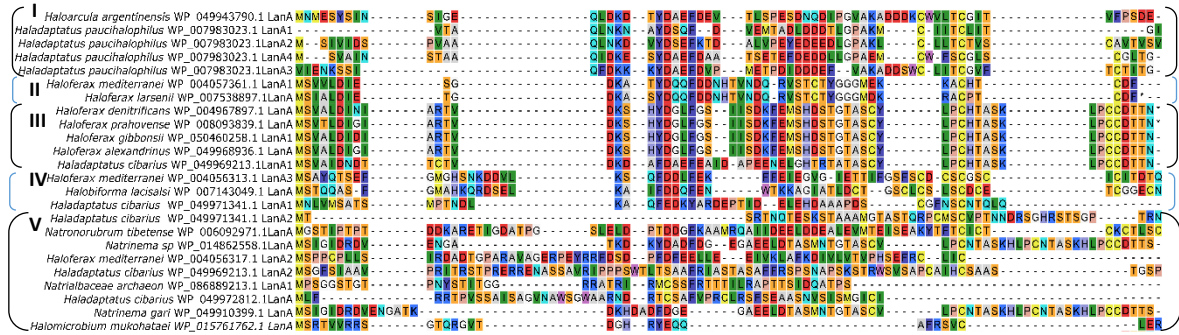


Figure 11 Multiple Sequence Alignment of all manually identified putative precursor peptides. In brackets are the four distinct groups identified.

Peptide alignment analysis showed that five distinct groups are formed (Figure 11), numbered from I to V. The first group is composed by four peptides identified in *Haladaptatus paucihalophilus* DX253 and the precursor peptide identified in the vicinity of *Haloarcula argentinensis* DSM 12282 LanM. The second group consists of peptides identified in two different species of *Haloferax* (*Haloferax larsenii* JCM 13917 and *Haloferax mediterranei* ATCC 33500). The third group is composed by peptides identified in five different Archaea. The fourth group has peptides identified in *Halobiforma lacisalsi* AJ5 and again in *Haloferax mediterranei* ATCC 33500 and *Haladaptatus cibarius* D43. The *H. mediterranei* ATCC 33500 peptide is encoded in a plasmid. The last group is composed by peptides from different genus that do not share similarity.

As abovementioned, class II lanthipeptides have some conserved motifs. One of these motifs is the so-called double-Gly motif (GG/GA/GS) and is located at the end of the leader sequence. It is recognized by the N-terminal papain-like C39 protease of the SunT-type LanT that removes the leader sequence and exports the class II lanthipeptides outside of the cells. This type of sequence is present in two peptides of the second group and in almost all the peptides of the third group (Figure 11). The other conserved motif of class II peptides is ELXXBX, which is essential for peptide recognition and correct modification by the LanM enzymes^{7,8}. This motif is not present in Archaeal precursor peptides, suggesting that archaeal LanM enzymes may not require these amino acids to recognize their cognate peptides.

In order to better understand the four groups sharing some homology, they were independently analyzed (Figure 12).

In the first group of Figure 12, several residues located in the N-terminus region of the precursor peptide were conserved constituting a QxxKxxYDxxF motif. Towards their C-terminus, there is also a highly conserved AK motif. It can be suggested that this motifs might be responsible for peptide recognition and peptidase activity as seen with the common class II ELXXBX and double-Gly motifs. Another interesting observation is the fact that several homologous peptides are present in the vicinity of *Haladaptatus paucihalophilus* DX253 LanM. This indicates that this enzyme has a high substrate tolerance and can modify, at least, four peptides. It is worth to notice that the C-terminus of these peptides has few conserved residues. Some LanM enzymes are known to modify different peptides with similar leader sequences. The best characterized example is ProcM that modify up to 29 different substrates with diverse ring topologies⁷³.

The second group (Figure 12) is composed by peptides that are very similar and that will also originate very similar mature peptides, regardless of being modified by LanM proteins from different species. This should imply that their cognate LanM enzymes are very similar as well.

The third group (Figure 12) also contains peptides encoded in the genome of five different species, but that share very high sequence homology. The most dissimilar peptide is from *Haladaptatus cibarius* D43. Even so, all of the organisms have the potential to produce the same mature peptide with the sequence TASCYLPCHTASKLPCCDTTN, suggesting that proteolysis can occur after the TG/TA motif and not after the GG motif abovementioned. In the N-terminus region is possible to identify a TVDK motif that can be responsible for peptide recognition. The conservation of the residues D, F, I and E might also indicate a possible function in export and/or transport.

In the fourth group (Figure 12) the motif LKXXFXD is conserved in the N-terminal region that might also have a recognition or cleavage function.

Normally, very identical biosynthetic enzymes modify very identical peptides. Moreover, the biosynthesis of closely-related peptides is expected to be similar. Therefore, the comparison of the genetic environment of *lanA* and *lanM* genes from these different groups can be useful for the identification of the enzymes that can be involved in the biosynthesis of lanthipeptides in Archaea. With that in mind, identification and comparison of putative clusters using Bagel software⁴⁸ followed.



Figure 12 Multiple sequence alignment between the four groups of similar putative LanA peptides.

3.3.1 Identification of gene clusters encoding the biosynthesis of Class II lanthipeptides in Euryarchaeota

We used Bagel⁴⁸, one of the many automated genome mining software available, to define the constitution of the possible lanthipeptide clusters (Figure 13).

The clusters were arranged in four different groups based on the results of the putative peptides alignment (Figure 13). A first analysis, shows that they are somehow different from the known, classical class II clusters. For instance, there are no genes encoding the bifunctional SunT-type LanT, responsible for the transport and maturation of the class II peptides. Nevertheless, we identified several ABC transporter-related proteins that can be involved either in transport or peptide immunity.

Different genes compose the gene clusters identified by Bagel. However, this does not imply that all of these genes are, in fact, part of the lanthipeptides biosynthetic cluster.

In the first group, the upstream and downstream genes of *lanM* are distinct. Therefore, based on the genetic context, it is only possible to identify a *lanM* and *lanA* gene as part of these clusters (Figure 13).

In the second group, the gene immediately upstream of *lanM* encodes a protein of unknown function (UF; Figure 13), which is likely involved in the biosynthesis of the lanthipeptide. The other genes encode metabolic proteins that are also present in Halobacteria genomes without LanMs and therefore should not be exclusively involved in the biosynthesis of lanthipeptides.

Focusing on the third group, it is worth to mention that Bagel failed to retrieve a cluster from the *Haloferax alexandrinus* JCM 10717 genome. The clusters identified for this group were very similar, except the one of *H. cibarius*. The clusters with intra-similarity (*H. denitrificans*, *H. prahovense* and *H. gibbonsii*) encode an ABC transporter immediately upstream of *lanM* genes. In *H. cibarius* D43, a homologue of this ABC transporter is also found upstream of its *lanM* gene,

but is not depicted in the Figure 13 because Bagel did not consider it. Thus, it is highly probable that the biosynthetic clusters of peptides from group three, are composed, at least, by a *lanM*, a *lanA* and two genes encoding an ABC transporter (one encodes the ATP-binding cassette and the other the transmembrane domain). The other genes that are very similar in *H. denitrificans*, *H. prahovense* and *H. gibbonsii* clusters are not present in the surroundings of *H. cibarius lanM*. This suggests that they must be not involved in the biosynthesis of the lanthipeptides.

In group four, apart from *lanM* and *lanA*, it seems that the three clusters do not share similarity among each other. Therefore, as for group 1, it is difficult to infer about the lanthipeptide cluster based on similarity. Nevertheless, it is worth to mention that the *lanM* gene of *H. lacisalsi* AJ5 is embedded in two ABC transporters (Bagel did not identify the permease ORF of one of the ABC transporter). Curiously, the gene immediately downstream of the second ABC transporter and a gene located some ORFs upstream of *lanM* encode transposases. This can suggest that all of the genes between the transposases can be important for the biosynthesis of the lanthipeptide. The *H. mediterranei* cluster will be discussed in more detail in the following section.

Some of the *lanA* genes identified by Bagel did not correspond to the peptides that were detected by “manual” search and described in the previous section. One limitation of Bagel software is the use of a particular set of characteristics based on the current knowledge pertaining a natural product and its biosynthetic machinery. In some cases, the software did not predict a cluster, in others it was unable to identify appropriate *lanA* genes or components of the ABC transporters. Moreover it predicts big biosynthetic clusters that very often contain genes that are ubiquitous in Archaea and that, therefore, are not directly related to lanthipeptide biosynthesis. These are also disadvantages of other prediction softwares, such as AntiSMASH. Therefore, the automatic analysis should be always carefully revised when it is intended to characterize the biosynthetic clusters of lanthipeptides (as well as other secondary metabolites). Finally, lanthipeptide enzymatic machinery usually cluster together in the genome. Nonetheless, proteins involved in the biosynthetic process can also be present somewhere in the genome.

3.4 The case study of *Haloferax mediterranei* ATCC 33500

Haloferax mediterranei ATCC 33500 received further attention for several reasons. Firstly, it encodes a total of three LanMs in its genome, both in the chromosome and in one of its plasmids (Figure 14). In the plasmid, the two *lanM* are contiguous, resembling two component peptides. Type II lanthipeptides are notorious for their two component peptides, where each precursor peptide is traditionally post translationally modified by a cognate enzyme. However, the bioactivity of these peptides depends on their synergy⁴.

In *Haloferax mediterranei* ATCC 33500, the genes adjacent to the *lanMs* were analysed in more detail (Figure 14 and Appendix 2). For the sake of simplicity, the LanMs genes were termed *medM1* for the one present in the chromosome cluster and *medM2* and *medM3* for those present in the plasmid. The putative peptides genes followed the same terminology and therefore were designated *medA1* (present in the chromosome) and *medA2* and *medA3* (in the plasmid).

In the chromosome, immediately upstream of the *lanA1* gene there is a gene encoding an acetate-CoA ligase family protein (HFX_RS04215). This enzyme is part of the core genome of various Halobacteria. Thus, *lanA* establishes the upstream limit of the biosynthetic cluster. Immediately downstream of *medM1* there is a gene that encodes a hypothetical protein (HFX_RS04205) that is not found in other microorganisms and that is probably transcribed together with *medM1*. Further downstream we identified two genes encoding proteins and an ABC transporter (HFX_RS04200 to HFX_RS04185) that have highly conserved homologues in most of the genomes of Halobacteria species. For this reason, these genes should not be involved in the biosynthesis of the *H. mediterranei* lanthipeptide encoded in the chromosome.

In the plasmid, immediately upstream of the *medM2* there are three genes (HFX_RS15975 to HFX_RS15985) that encode proteins with high similarity with proteins that are also encoded in the genomes of Halobacteria that lack *lanM* genes. The same was observed for the oleate hydratase (HFX_RS15995) encoded between *medM2* and *medM3* genes. For this reason, it is probable that these genes are not essential for the biosynthesis of MedA2 and MedA3 peptides. Immediately downstream of the *medM3* gene there are two genes encoding an ABC transporter (HFX_RS16010 and HFX_RS16015) and a gene encoding a hypothetical protein (HFX_RS16020). ABC-transporters are a family of integral membrane proteins transporters widely spread among organisms. They are characterized by their nucleotide-binding and transmembrane domains⁷⁴. However, very frequently the transmembrane domains are annotated as hypothetical proteins. In order to confirm the presence of transmembrane domains in HFX_RS16010, TMHMM was used for further analysis. TMHMM is a membrane protein topology prediction method based on a hidden Markov

model. It predicts transmembrane helices in proteins and allows distinguish between soluble and membrane bound proteins with a high degree of accuracy. The analysis (Figure 14) confirmed that the protein encoded by HFX_RS16010 has 6 transmembrane domains. Genes encoding a similar ABC transporter and a similar hypothetical protein were identified in *Halobacteria* containing *lanM* genes. Moreover, they do not seem to be ubiquitous in *Halobacteria* species. This indicates that these genes are involved in the biosynthesis of either MedA3 or MedA2 or even both. The *nisFEG* locus in nisin biosynthetic cluster also have similar ABC transporters. NisF has an ATP-binding site domain and the transmembrane hydrophobic proteins NisE/G work together for peptide excretion from the cell. Therefore, it is not clear if the ABC transporter identified in the plasmid biosynthetic cluster can be involved in a putative transport of the lanthipeptides from the cells or if, alternatively, they are a self-immunity system. The last hypothesis imply that the lanthipeptides encoded in the plasmid have archeocin activity.

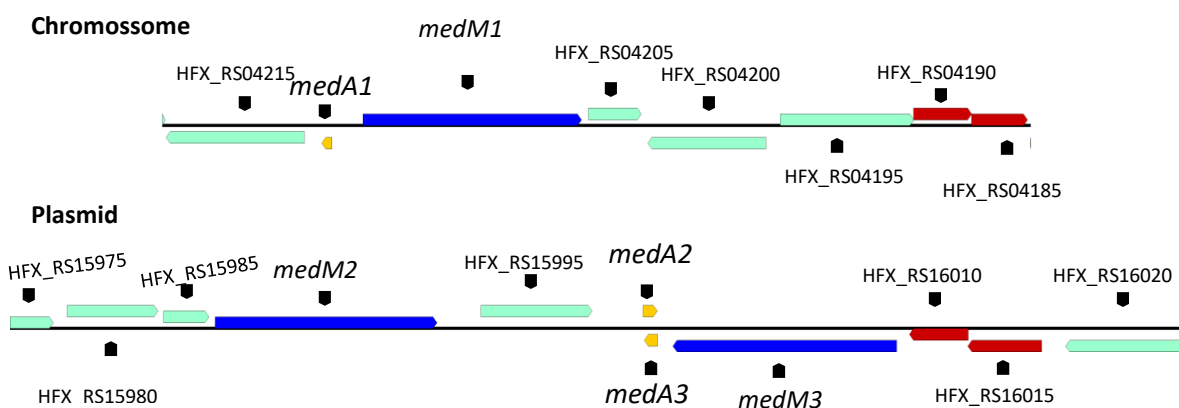


Figure 14 Schematic representation of the two clusters identified in *Haloferax mediterranei*. In blue the LanMs genes identified. In yellow the putative peptides genes. In red the ABC transporter genes.

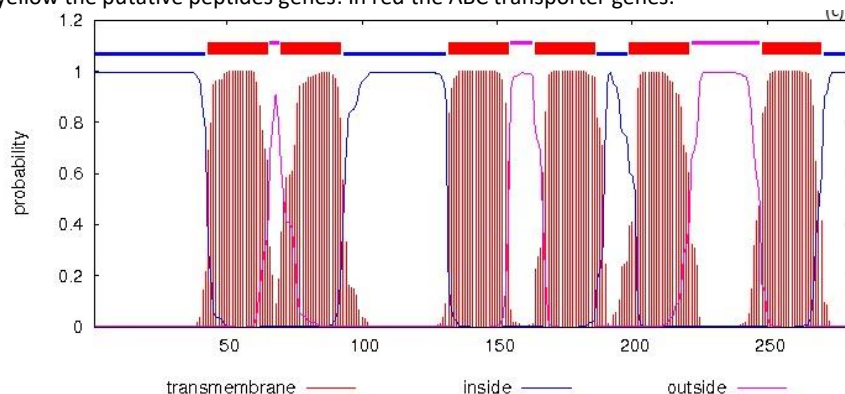


Figure 15 Plot generated by TMHMM analysis of the gene *HFX_RS16010* coding for the transmembrane domain of an ABC transporter.

3.4.1 *Escherichia coli* co-expression of the putative *medA* and *medM* genes from *Haloferax mediterranei* ATCC 33500

E. coli is a well-established model organism for lanthipeptide production. Several studies of heterologous co-expression revealed paramount for the identification of several lanthipeptides and to understand their respective clusters. With that in mind, the *E. coli* model seemed appropriate for the co-expression of the biosynthetic enzymes identified in *H. mediterranei* and their respective putative precursor peptides. To that end, the pRSFDuet-1 plasmid was chosen as the expression vector for this study. One of the multi cloning sites (MCS) was used to house the precursor peptide gene (*medA1/medA2/medA3*), while the other MCS housed the corresponding biosynthetic enzyme (*medM1/medM2/medM3*). The construction was performed in order to express a His₆Tag fused to the N-terminal of MedAs peptides. This would allow recovering the modified peptides by IMAC. The role of this enzyme is the dehydration of the serine and threonine residues, followed by cyclization from the attack of cysteine. This dehydration would mean an overall decrease in the peptide molecular weight. With each water molecule lost, a decrease in 18Da is expected. The serine and threonine residues located in the N-terminal region of the peptide, usually associated with the leader sequence, are not expected to be involved in the dehydration reaction. But, as no clear distinction between leader sequence and core sequence is possible at the moment, all serine and threonine residues present in the precursor peptide are considered putative targets for the reaction (marked in red in Figure 16). Based on this, it is possible to predict the masses of the different peptides that can result after dehydration by MedMs (Appendix 3). When we further analyzed the MedA C-terminus, it is also possible to predict that MedA1 has the potential to form three Lan/MeLan rings, MedA2 can form two and MedA3 five, based on their number of cysteines (Figure 16).

```
RSFMedA1  GSSHHHHHSQDPSVVLDIESGDKATYDQQFDDNHTVNDQRVSTCTYGGGMEKKACHTCDF
RSFMedA2  GSSHHHHHSQDPSPPCPLLSIRDADTGPARAVAGERPEYRRFDSDPDFEELLEIVKLAFKDIVLVTVPHSEFRCLIC
RSFMedA3  GSSHHHHHSQDPSAYQTSEFFGMGHSNKDDVLKSQFDDLFEKFFEIEGVGITTIFGSFSCDCSCGSCICITDTQ
```

Figure 16 Sequence of the His₆-tagged precursor peptides for *E. coli* co-expression. Colored in red are the serine and threonine residues essential for the dehydration reaction.

After co-expression of *medA* and *medM* genes and purification by standard methods, the MedA peptides were analyzed by mass spectrometry (MS). It was not possible to identify MedA masses shift corresponding to dehydration reactions, indicating that MedM enzymes were not active. Confirmation of MedM expression after IPTG induction under various conditions was investigated by SDS-PAGE of whole cell protein extracts (Figure 17). We confirmed that MedM proteins were

expressed, but they are inactive, most probably due to the lack of correct folding in the *E. coli* expression system. Being an Archaeal gene, the transcription of our biosynthetic enzyme can be compromised simply due to codon preference of the transcriptional machinery of *E. coli*. Certain archaeal organisms prefer codons recognized by tRNAs that are rarely used in *E. coli*, making transcription and proper folding difficult in this model⁷⁵. Another known issue of this bacterial model is the formation of inclusion bodies meaning that the translated protein ends up an agglomerate of precipitated peptides with no utility whatsoever. Moreover, given that the MedMs could have evolved to be functional in a high salinity environment, it is possible that they misfold in the *E. coli* cytoplasm. To overcome this issues, *in vitro* experiments have been used to validate PTM activity of multiple LanM enzymes^{76,77}. This approach could be employed to investigate the MedMs activity on MedAs. But given the existence of many variables that can not be easily addressed, such as ensuring the proper folding of biosynthetic enzymes of our isolates, it was decided to change the model organism for co-expression of the genes under study and use instead the archaea *Haloferax volcanii* H1424.

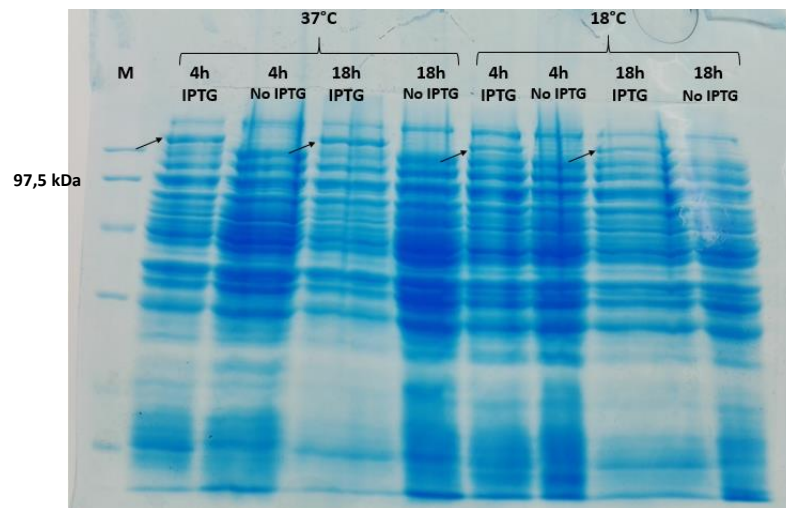


Figure17 SDS-Page of whole cell proteins. The cells harboring the gene *medM3* were used for protein expression confirmation. Results from 4 and 18 hours incubation at 37°C and 18°C after induction of expression by IPTG confirmed expression of our 120 kDa enzyme (arrows). Respective negative controls are present.

3.4.2 *Haloferax Volcanii* H1424 co-expression of the putative *medA* and *medM* genes from *Haloferax mediterranei* ATCC 33500

Haloferax volcanii, isolated from the Dead Sea, is the quintessential model organism for Archaeal genetics. *Haloferax volcanii* is highly similar to *Haloferax mediterranei*, as confirmed from genome comparison⁶⁴. The use of similar organisms for heterologous expression should mitigate some of the common issues encountered with the *E. coli* model. Several strategies have been applied to engineer *H. volcanii* strain for genetics studies and heterologous gene expression purposes. The H1424 strain is the culmination of several efforts that worked towards that end. The vector created for the purpose of protein expression was plasmid pTA1392 and it was used in the present work. This plasmid has an inducible tryptophanase promoter (p.tnaA) and a pHV2 origin of replication that maintains the plasmid at a copy number of approximately ~6 per genome equivalent⁵⁷. Overexpression is induced by the addition of tryptophan to the culture medium. An ampicillin resistance gene and a high copy number origin of replication allow for the use of the classical methodologies for cloning and plasmid amplification in *E. coli*. However, pTA1392 has a single promoter. Therefore we developed a strategy to use this plasmid for co-expression of *medA* and *medM* genes where each of these genes is under the control of a p.tnaA promoter. This system is described in the following section.

3.4.2.1 Plasmid pTA1392 construction for co-expression

Using *E. coli* as the heterologous expression system for lanthipeptides has proved invaluable in the study of this RiPP family. From expression of enzymes for *in vitro* studies^{76,77} to *in vivo* reconstitution of a complete lanthipeptide biosynthetic pathway in a single vector², *E. coli* has been extensively used. The co-expression of an enzyme and its substrate in a single vector is a very elegant way of studying protein activity. Archaeal genetics tools and expression systems have lagged behind compared to bacteria and eukaryotes. Nevertheless, it is available today the highly engineered *H. volcanii* H1424 strain and the corresponding pTA1392 plasmid designed for inducible protein expression. As above-mentioned, this vector has a single multiple cloning site with a single promoter. Mimicking the pRSFDuet-1 design, an attempt was made to produce pTA1392 plasmids containing *medA* and *medM* genes under the control of two tryptophan promoters, respectively. These plasmids were constructed in a two-step fashion (Figure 18). Firstly, the genes of interest (all *medA* and *medM* genes) were individually inserted in the MCS of

pTA1392. The genes for the putative precursor peptides were cloned by allowing for the expression of a His₆-tag on their N-terminus. The *medM* genes were cloned in order to contain no tags. In the second step, the plasmids containing *medA* genes were used to amplify the region containing the p.tnaA promoter and the His₆-tag fused *medA*. This fragment was further cloned in the downstream region of *medM* genes already cloned in pTA1392.

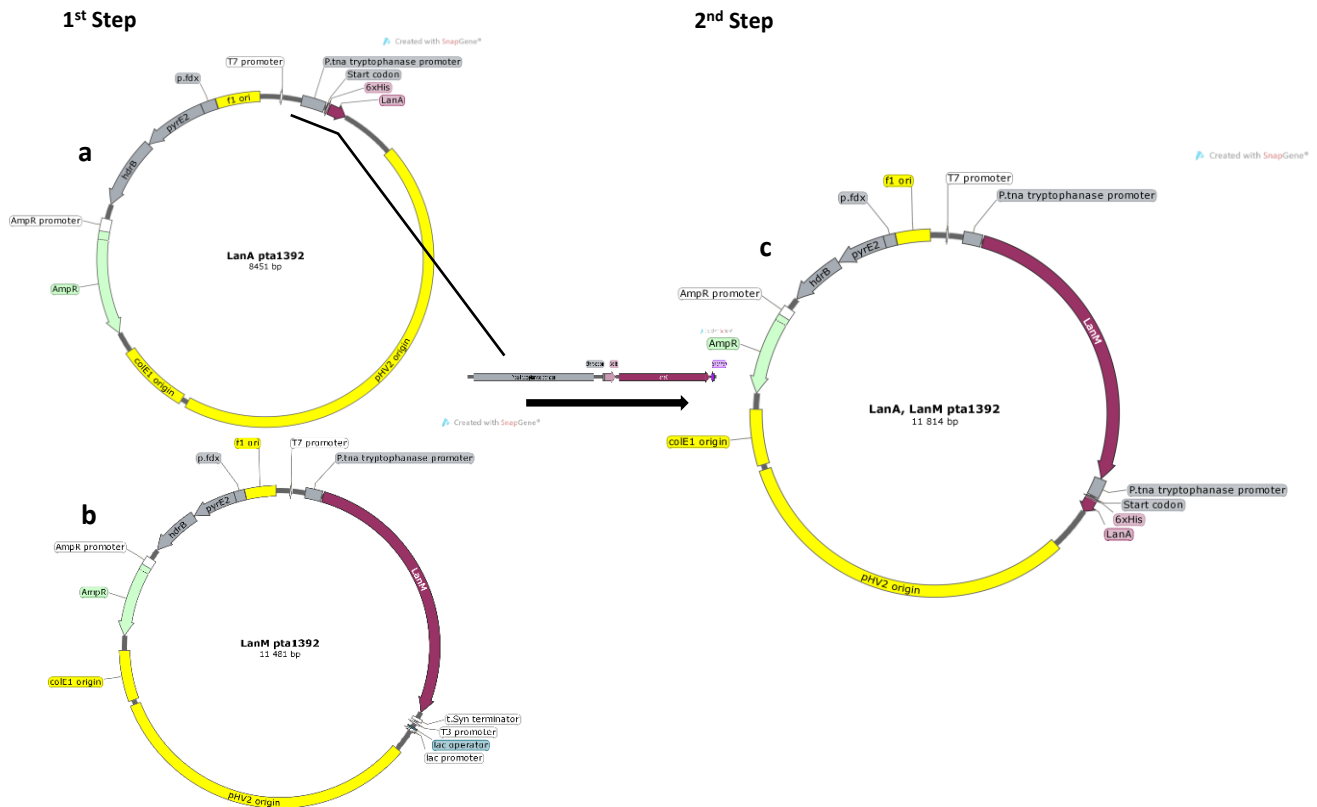


Figure 18 Schematics of plasmid construction approach used for co-expression of *medA* and *medM* genes in *Haloferax volcanii*. a) vector pTA1392 with *medA* gene, b) vector pTA1392 with *medM* gene, *) sequence of the promoter and the *medA* gene fused to the His₆-Tag, c) final construct containing both genes associated with their p.tnaA promoters.

3.4.2.2 *Haloferax volcanii* H1424 protein expression results

The design of the pTA1392 for co-expression introduced some modifications to the vector. Therefore, to assure that the system was working before these modifications were introduced, we first expressed and purified only His₆-tagged MedA peptides from *H. volcanii* containing pTA1392 with *medA* genes (Figure 19). The expected average molecular masses for His₆-MedA1, His₆-MedA2 and His₆-MedA3 were 6470.987 Da, 8731.93 Da and 7960.749 Da, respectively. These masses were calculated based on the fact that the first Met of the peptide is removed after

ribosomal synthesis. This was because the initiator Met is removed in approximately two-thirds of the haloarchaeal proteins⁷⁸. The IMAC eluted peptides samples were analyzed directly (after ziptip desalting) and after SPE by MS. Unfortunately, it was not possible to detect masses within the expected molecular weights (Figure 20). This also applies if we do not consider the N-terminal removal of Met.

pTAMedA1 HHHHHHMMW^SVVLDIE^SGDKA^TYDQQFDDNHTVNDQQRV^STCTYGGGMEKKACHTCDF
pTAMedA2 HHHHHHMMW^SPPCP^LLSIRDAD^TGPARAVAGERPEYRRFD^SDPFD^SFEELLEIVKLA^FKDIVLV^TVPH^SEF^RCLIC
pTAMedA3 HHHHHHMMW^SAYQT^SSEFGMGH^SNKDDVLK^SQFDDLFEKFFEIEGVGI^TTTIFGS^SFCDC^SCG^SSCICITDTQ

Figure 19 Sequence of the His₆-tagged precursor peptides for *H. volcanii* co-expression. Colored in red are the serine and threonine residues essential for the dehydration reaction.

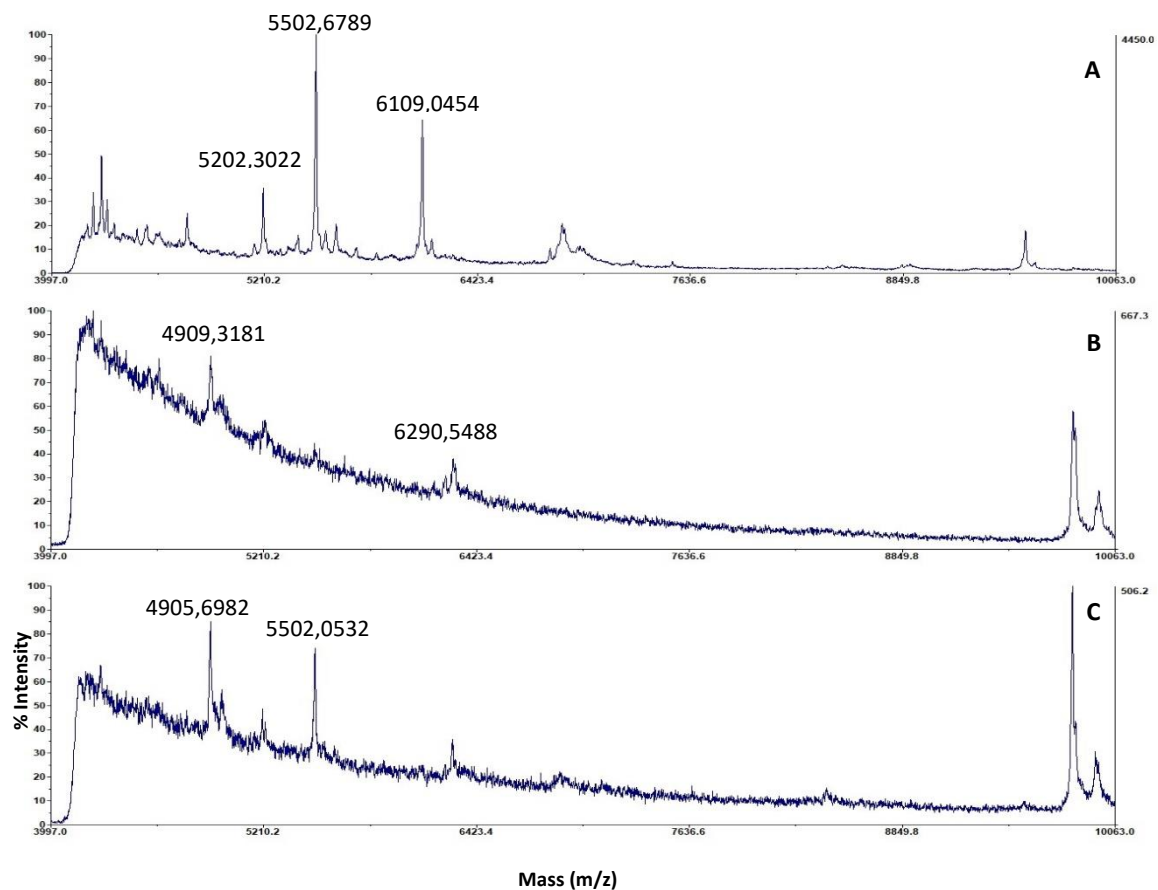


Figure 20 Mass spectra of *Haloferax volcanii* expression products after IMAC purification and ZipTip desalting. A) *Haloferax volcanii* transformed with plasmid containing the chromosomal *medA1*; B) *Haloferax volcanii* transformed with plasmid containing *medA2*; C) *Haloferax volcanii* transformed with plasmid containing *medA3*.

The immediate conclusion is that either the protein expression or the protein purification protocols must be optimized. His₆-tagged LanA peptides are typically overexpressed insolubly in the *E. coli* model⁶⁸. The use of the strong denaturant agent guanidine hydrochloride became

therefore standard practice in purification protocols of lanthipeptides produced in *E. coli*. This agent could be included in further refinement and optimization of peptide purification procedures produced in *H. volcanii*, which will be essential for future studies of archaeal lanthipeptides.

Other hypothesis is that production of MedAs is achieved but the linear peptides are cleaved by a non-specific *H. volcanii* peptidase. Given that no obvious peptidase is encoded in the *H. mediterranei* ATCC 33500 clusters, leader peptide removal could be performed by a protease that is elsewhere in its genome. This is observed in the biosynthesis of Class III and IV lanthipeptides, where the leader sequences are cleaved off, after peptide modification, by an unknown peptidase that is not encoded in the cluster. The result of this proteolysis are mature peptides of varying lengths²⁹. *H. mediterranei* and *H. volcanii* chromosomal genomes share high homology⁶⁴. Consequently, *H. volcanii* can encode a protease that targets the leader sequences of MedA peptides, even if it does not have lanthipeptide clusters. Under such circumstances, it would be only possible to purify the leader peptides of His₆-MedAs.

4. Final Remarks

The natural world is still the major source of novel and exciting molecules. The use of bioinformatics tools in this genomic era will certainly continue to prove extremely useful in discovering this natural wonders of ribossomally synthesized and post transnationally modified peptides.

Lanthipeptides, which were first thought to have mainly antibiotic activity are now endowed with a wider spectrum of bioactivities. Continuous efforts in genome mining will certainly uncover more. With this work, we aimed at identifying and characterizing, for the first time, lanthipeptides and the corresponding lanthipeptide modifying enzymes, LanMs, from the Domain Archaea. These LanM enzymes showed high homology to those present in the Bacteria and contain all the residues essential for enzyme function. Comparison of the putative precursor lanthipeptides encoded in different archaeal organisms revealed the conservation of promising, characteristic motifs and residues that might be essential for peptide-enzyme interactions. Analysis of the clusters with Bagel revealed little similarity with classical class II clusters lacking, for instance, the SunT-type LanT genes coding for bi-functional transporter proteins. These proteins are responsible for the final stages of lanthipeptide maturation, consisting on their cleavage and excretion from the cell. Transcription regulation genes were also lacking in most of the clusters. However, the presence of ABC transporters was common and can be associated either with export or self-immunity. Self-immunity will be relevant only if these lanthipeptides are archaeocins.

Focusing on *H. mediterranei* ATCC 33500, it was found that the chromosomal cluster is most probably composed by *lanA*, *lanM* and also by a gene encoding a protein of unknown function. Regarding the plasmid cluster, it should contain *medM2*, *medA2*, *medA3* and *medM3* genes together with genes encoding an ABC transporter and a gene encoding a protein of unknown function. However, the confirmation of lanthipeptide presence will be required before function of the unknown proteins can be further investigated.

The putative lanthipeptides explored in this study do not share homology with classical class II peptides. New peptides possibly encode novel bioactivity. If Archaea lanthipeptides have archaeocin activity, they would prove a valuable tool, for instance, in the development of Archaeal genetics tool-box. Archaeal genetic systems have always lagged behind the more sophisticated and extensively used bacterial and eukaryotic ones like those of *E. coli* or *Saccharomyces cerevisiae*. This was in part due to the lack of ready to use selective markers and their resistance genes. Most of the bacterial antibiotics that are readily available for the development of bacterial genetic manipulation systems, are not usable in archaeal systems.

Therefore, in archaea, the selection is mainly based on auxotrophic strains. If the archaeal lanthipeptides identified in this study proved to have archaeocin activity and a self-immunity system (conferred by ABC transporters), they could be used in archaea to substitute the need for auxotrophic strains. This would obviously imply the need for bioactivity assays after archaeal lanthipeptide identification and characterization.

Not only are the lanthipeptides interesting, as are the biosynthetic enzymes responsible for their post-translational modifications. Control over stereoisomery is one of the greatest challenges on a chemistry laboratory. The characteristic spatial conformations of class II peptides is mainly managed by LanMs. Substrate promiscuity was also confirmed in some enzymes. The possibility of producing chimeras consisting of the leader peptide sequence recognizable by an enzyme capable of a specific conformational modification of interest and a core peptide of our choice that can then be subjected to these new LanM catalyzed modifications, has an enormous potential for biotechnological applications. Recently, a protocol for a *H. volcanii* expression system in industrial settings was developed to tackle the environmental limitations of the preferred models for biotechnological applications, like the *E. coli* model. The fact that these enzymes evolved to function in high salinity and temperature environments, typical conditions of these industrial bioreactors, makes it sound to assume these archaeal lanthipeptide modifying enzymes extremely interesting for those same biotechnological applications at an industrial scale.

The discovery of novel natural products is largely based on trial and error, where the need for improved tools is constant. With this work we constructed a vector that can be used for co-expression of *medA* and *medM* genes in Archaea. This protocol could have tremendous applications for future studies in archaeal biosynthetic machinery and novel lanthipeptide discovery. Unfortunately, we were unable to establish the proof of concept that MedM enzymes do modify MedA peptides. In a near future, the protocols necessary for their co-expression and purification will be optimized in order to confirm that *Haloferax mediterranei* ATCC 33500 is a lanthipeptide producer. After this, the concept will have the potential to be applied to other archaeal lanthipeptides or RiPPs in general.

5. Bibliography

1. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **79**, 629–661 (2016).
2. Caetano, T., Krawczyk, J. M., Mösker, E., Süßmuth, R. D. & Mendo, S. Heterologous expression, biosynthesis, and mutagenesis of type II lantibiotics from *Bacillus licheniformis* in *Escherichia coli*. *Chem. Biol.* **18**, 90–100 (2011).
3. McIntosh, J. A., Donia, M. S. & Schmidt, E. W. Ribosomal peptide natural products: bridging the ribosomal and n1. McIntosh JA, Donia MS, Schmidt EW. Ribosomal peptide natural products: bridging the ribosomal and nonribosomal worlds. *Nat. Prod. Rep.* **26**, 537–59 (2009).
4. Bierbaum, G. & Sahl, H.-G. Lantibiotics: mode of action, biosynthesis and bioengineering. *Curr. Pharm. Biotechnol.* **10**, 2–18 (2009).
5. Petrova, M. I. *et al.* The Lantibiotic Peptide Labyrinthopeptin A1 Demonstrates Broad Anti-HIV and Anti-HSV Activity with Potential for Microbicidal Applications. *PLoS One* **8**, (2013).
6. Ortega, M. A. & van der Donk, W. A. New Insights into the Biosynthetic Logic of Ribosomally Synthesized and Post-translationally Modified Peptide Natural Products. *Cell Chem. Biol.* **23**, 31–44 (2016).
7. Repka, L. M., Chekan, J. R., Nair, S. K. & Donk, W. A. Van Der. Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes. *Chem. Rev.* **117**, 5457–5520 (2017).
8. Yu, Y., Zhang, Q. & van der Donk, W. A. Insights into the evolution of lanthipeptide biosynthesis. *Protein Sci.* **22**, 1478–89 (2013).
9. Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30**, 108–60 (2013).
10. Schnell, N. *et al.* Prepeptide sequence of epidermin, a ribosomally synthesized antibiotic with four sulphide-rings. *Nature* **333**, 276–278 (1988).
11. Park, S. & James, C. D. Lanthionine Synthetase Components C-like 2 Increases Cellular Sensitivity to Adriamycin by Decreasing the Expression of P-Glycoprotein through a transcription-mediated Mechanism. *Cancer Res.* **63**, 723–727 (2003).
12. Zeng, M., Donk, W. A. Van Der & Chen, J. Lanthionine synthetase C – like protein 2 (LanCL2) is a novel regulator of Akt. *Mol. Biol. Cell* **2**, (2014).
13. Lubelski, J., Rink, R., Khusainov, R., Moll, G. N. & Kuipers, O. P. Biosynthesis, immunity, regulation, mode of action and engineering of the model lantibiotic nisin. *Cell. Mol. Life Sci.* **65**, 455–76 (2008).

14. Wiedemann, I. *et al.* Specific Binding of Nisin to the Peptidoglycan Precursor Lipid II Combines Pore Formation and Inhibition of Cell Wall Biosynthesis for Potent Antibiotic Activity. *J. Biol. Chem.* **276**, 1772–1779 (2001).
15. Hooven, H. W. Van Den *et al.* NMR and circular dichroism studies of the lantibiotic nisin in environments. *FEBS Lett.* **319**, 189–194 (1993).
16. Hsu, S. D. *et al.* The nisin – lipid II complex reveals a pyrophosphate cage that provides a blueprint for novel antibiotics. *Nat. Struct. Mol. Biol.* **11**, 963–967 (2004).
17. Zhang, Q., Doroghazi, J. R., Zhao, X., Walker, M. C. & van der Donk, W. A. Expanded Natural Product Diversity Revealed by Analysis of Lanthipeptide-Like Gene Clusters in Actinobacteria. *Appl. Environ. Microbiol.* **81**, 4339–4350 (2015).
18. Koponen, O., Takala, T. M., Saarela, U., Qiao, M. & Saris, P. E. J. Distribution of the NisI immunity protein and enhancement of nisin activity by the lipid-free NisI. *Microbiol. Lett.* **231**, 85–90 (2004).
19. Stein, T., Heinzmann, S., Solovieva, I. & Entian, K.-D. Function of *Lactococcus lactis* nisin immunity genes *nisl* and *nisFEG* after coordinated expression in the surrogate host *Bacillus subtilis*. *J. Biol. Chem.* **278**, 89–94 (2003).
20. Koponen, O. *et al.* NisB is required for the dehydration and NisC for the lanthionine formation in the post- translational modification of nisin. *Microbiology* **148**, 3561–3568 (2002).
21. Garg, N., Salazar-Ocampo, L. M. A. & van der Donk, W. A. In vitro activity of the nisin dehydratase NisB. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 7258–7263 (2013).
22. Ortega, M. A. *et al.* Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB. *Nature* **517**, 509–12 (2015).
23. Li, B. & van der Donk, W. A. Identification of essential catalytic residues of the cyclase NisC involved in the biosynthesis of nisin. *J. Biol. Chem.* **282**, 21169–75 (2007).
24. Tyne, D. Van, Martin, M. J. & Gilmore, M. S. Structure, Function, and Biology of the *Enterococcus faecalis* Cytolysin. *Toxins (Basel)*. **5**, 895–911 (2013).
25. Dong, S. *et al.* The enterococcal cytolysin synthetase has an unanticipated lipid kinase fold. *Elife* **4**, e07607 (2015).
26. Tang, W., Jiménez-osés, G., Houk, K. N. & Donk, W. A. Van Der. Substrate control in stereoselective lanthionine biosynthesis. *Nat. Chem.* **7**, 57–64 (2015).
27. Kodani, S. *et al.* The SapB morphogen is a lantibiotic-like peptide derived from the product of the developmental gene *ramS* in *Streptomyces coelicolor*. *Proc. Natl. Acad. Sci. U. S. A.*

- 101**, 11448–53 (2004).
28. Goto, Y. *et al.* Discovery of Unique Lanthionine Synthetases Reveals New Mechanistic and Evolutionary Insights. *PLoS Biol.* **8**, e1000339 (2010).
 29. Meindl, K. *et al.* Labyrinthopeptins: a new class of carbacyclic lantibiotics. *Angew. Chem. Int. Ed. Engl.* **49**, 1151–1154 (2010).
 30. Jungmann, N. A., Krawczyk, B., Tietzmann, M., Ensle, P. & Süssmuth, R. D. Dissecting Reactions of Nonlinear Precursor Peptide Processing of the Class III Lanthipeptide Curvopeptin. *J. Am. Chem. Soc.* **136**, 15222–15228 (2014).
 31. Oman, T. J., Knerr, P. J., Bindman, N. A., Vela, J. E. & Donk, W. A. Van Der. An Engineered Lantibiotic Synthetase That Does Not Require a Leader Peptide on Its Substrate. *J. Am. Chem. Soc.* **134**, 6952–6955 (2012).
 32. Müller, W. M., Ensle, P., Krawczyk, B. & Süssmuth, R. D. Leader Peptide-Directed Processing of Labyrinthopeptin A2 Precursor Peptide by the Modifying Enzyme LabKC. *Biochemistry* **50**, 8362–8373 (2011).
 33. Klein, C., Kaletta, C., Schnell, N. & Entian, K. D. Analysis of Genes Involved in Biosynthesis of the Lantibiotic Subtilin. *Appl. Environ. Microbiol.* **58**, 132–142 (1992).
 34. Complex, M. L. S., Siegers, K., Heinzmann, S. & Entian, K. Biosynthesis of Lantibiotic Nisin. *J. Biol. Chem.* **271**, 12294–12301 (1996).
 35. van der Meer, J. R. *et al.* Influence of Amino Acid Substitutions in the Nisin Leader Peptide on Biosynthesis and Secretion of Nisin by *Lactococcus lactis*. *J. Biol. Chem.* **269**, 3555–3562 (1994).
 36. Völler, G. H., Krawczyk, B., Ensle, P. & Süssmuth, R. D. Involvement and Unusual Substrate Specificity of a Prolyl Oligopeptidase in Class III Lanthipeptide Maturation. *J. Am. Chem. Soc.* **135**, 7426–7429 (2013).
 37. Allgaier, H., Jung, G., Werner, R. G., Schneider, U. & Zähler, H. Elucidation of the Structure of Epidermin, a Ribosomally Synthesized, Tetracyclic Heterodetic Polypeptide Antibiotic. *Angew. Chem. Int. Ed. Engl.* **24**, 1051–1053 (1985).
 38. Castiglione, F. *et al.* Determining the Structure and Mode of Action of Microbisporicin, a Potent Lantibiotic Active Against Multiresistant Pathogens. *Chem. Biol.* **15**, 22–31 (2008).
 39. Boakes, S., Cortés, J., Appleyard, A. N., Rudd, B. A. M. & Dawson, M. J. Organization of the genes encoding the biosynthesis of actagardine and engineering of a variant generation system. *Mol. Microbiol.* **72**, 1126–1136 (2009).
 40. Iorio, M. *et al.* A glycosylated, labionin-containing lanthipeptide with marked

- antinociceptive activity. *ACS Chem. Biol.* **9**, 398–404 (2014).
41. Siegers, K. D. Genes Involved in Immunity to the Lantibiotic Nisin Produced by *Lactococcus lactis* 6F3. *Appl. Environ. Microbiol.* **61**, 1082–1089 (1995).
 42. Aso, Y. *et al.* A Novel Type of Immunity Protein, NukH, for the Lantibiotic Nukacin ISK-1 Produced by *Staphylococcus warneri* ISK-1. *Biosci. Biotechnol. Biochem.* **69**, 1403–1410 (2005).
 43. Rourke, S. O., Widdick, D. & Bibb, M. A novel mechanism of immunity controls the onset of cinnamycin biosynthesis in *Streptomyces cinnamomeus* DSM 40646. *J. Ind. Microbiol. Biotechnol.* **44**, 563–572 (2017).
 44. Ruyter, P. G. G. A. D. E., Kuipers, O. P. & Beerthuyzen, M. M. Functional Analysis of Promoters in the Nisin Gene Cluster of *Lactococcus lactis*. *J. Bacteriol.* **178**, 3434–3439 (1996).
 45. McAuliffe, O., Keefe, T. O., Hill, C. & Ross, R. P. Regulation of immunity to the two-component lantibiotic, lactacin 3147, by the transcriptional repressor LtnR. *Mol. Microbiol.* **39**, 982–993 (2001).
 46. Schmitter, T., Wiedemann, I., Sahl, H. & Bierbaum, G. Role of the Single Regulator MrsR1 and the Two-Component System MrsR2 / K2 in the Regulation of Mersacidin Production and Immunity. *Appl. Environ. Microbiol.* **68**, 106–113 (2002).
 47. Hetrick, K. J., Donk, W. A. Van Der & Der, L. Ribosomally synthesized and post-translationally modified peptide natural product discovery in the genomic era. *Curr. Opin. Chem. Biol.* **38**, 36–44 (2017).
 48. Jong, A. De, Heel, A. J. Van, Kok, J. & Kuipers, O. P. BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Res.* **38**, 647–651 (2010).
 49. Weber, T. *et al.* antiSMASH 3.0 — a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43**, 237–243 (2017).
 50. Alkhatib, Z., Abts, A., Mavaro, A., Schmitt, L. & Smits, S. H. J. Lantibiotics: How do producers become self-protected? *J. Biotechnol.* **159**, 145–154 (2012).
 51. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiology* **5**, 1–17 (2014).
 52. Nagao, J. *et al.* Lanthionine introduction into nukacin ISK-1 prepeptide by co-expression with modification enzyme NukM in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* **336**, 507–513 (2005).
 53. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary

- kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5088–5090 (1977).
54. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms : Proposal for the domains. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 4576–4579 (1990).
 55. Allers, T. & Mevarech, M. ARCHAEL GENETICS — THE THIRD WAY. *Nature* **6**, 58–73 (2005).
 56. Bergerat, A. *et al.* An atypical topoisomerase II from Archaea with implications for meiotic recombination. *Nature* **386**, 414–417 (1997).
 57. Allers, T. Overexpression and purification of halophilic proteins in *Haloferax volcanii*. *Bioeng. Bugs* **1**, 288–290 (2010).
 58. Connaris, H., Chaudhuri, J. B., Danson, M. J. & Hough, D. W. Expression, reactivation and purification of Enzymes from *Haloferax volcanii* in *Escherichia coli*. *Biotechnol. Bioeng.* **64**, 38–45 (1999).
 59. Allers, T., Ngo, H., Mevarech, M. & Lloyd, R. G. Development of Additional Selectable Markers for the Halophilic Archaeon *Haloferax volcanii* Based on the *leuB* and *trpA* Genes. *Appl. Environ. Microbiol.* **70**, 943–953 (2004).
 60. Bitan-banin, G., Ortenberg, R. & Mevarech, M. Development of a Gene Knockout System for the Halophilic Archaeon *Haloferax volcanii* by Use of the *pyrE* Gene. *J. Bacteriol.* **185**, 772–778 (2003).
 61. Allers, T., Barak, S., Liddell, S., Wardell, K. & Mevarech, M. Improved Strains and Plasmid Vectors for Conditional Overexpression of His-Tagged Proteins in *Haloferax volcanii*. *Appl. Environ. Microbiol.* **76**, 1759–1769 (2010).
 62. Strillinger, E., Grötzinger, S. W., Allers, T., Eppinger, J. & Weuster-botsch, D. Production of halophilic proteins using *Haloferax volcanii* H1895 in a stirred-tank bioreactor. *Appl. Microbiol. Biotechnol.* **100**, 1183–1195 (2016).
 63. Rodriguez-Valera, F., Ruiz-Berraquero, F., Ramos-Cormenzana, A. Isolation of Extremely Halophilic Bacteria Able to Grow in Defined Inorganic Media with Single Carbon Sources. *J. Gen. Microbiol.* **119**, 535–538 (1980).
 64. Han, J. *et al.* Complete Genome Sequence of the Metabolically Versatile Halophilic Archaeon *Haloferax mediterranei*, a Poly (3-Hydroxybutyrate- co -3-Hydroxyvalerate) Producer. *J. Bacteriol.* **194**, 4463–4464 (2012).
 65. Finn, R. D. *et al.* InterPro in 2017 — beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, 190–199 (2017).
 66. Gerlt, J. A. *et al.* Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for

- generating protein sequence similarity networks. *Biochim. Biophys. Acta - Proteins Proteomics* **1854**, 1019–1037 (2016).
67. Shannon, P. *et al.* Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498–2504 (2003).
 68. Li, B., Cooper, L. E. & Donk, W. A. Van Der. in *Methods in Enzymology* **458**, 533–558 (Elsevier Inc., 2009).
 69. Cline, S. W., Lam, R. L., Charlebois, R. L., Schalkwyk, L. C. & Doolittle, W. F. Transformation methods for halophilic archaebacteria. *Can. J. Microbiol.* **35**, 148–152 (1989).
 70. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
 71. Oren, A., Ventosa, A. & Kamekura, M. in *Bergey's Manual of Systematics of Archaea and Bacteria* 1–5 (2017). doi:10.1002/9781118960608.cbm00026.pub2.
 72. Tietz, J. I. *et al.* A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **13**, 470–478 (2017).
 73. Li, B. *et al.* Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10430–10435 (2010).
 74. Wilkens, S. Structure and mechanism of ABC transporters. *F1000Prime Rep.* **7**, 1–9 (2015).
 75. Kim, S. & Lee, S. B. Rare codon clusters at 5' -end influence heterologous expression of archaeal gene in Escherichia coli. *Protein Expr. Purificait.* **50**, 49–57 (2006).
 76. Chatterjee, C., Patton, G. C., Cooper, L., Paul, M. & van der Donk, W. A. Engineering Dehydro Amino Acids and Thioethers into Peptides Using Lactacin 481 Synthetase. *Chem. Biol.* **13**, 1109–1117 (2006).
 77. Cooper, L. E., Fogle, E. J. & van der Donk, W. A. Nine Post-translational Modifications during the Biosynthesis of Cinnamycin. *J. am* **133**, 13753–13760 (2011).
 78. Falb, M. *et al.* Archaeal N-terminal Protein Maturation Commonly Involves N-terminal Acetylation : A Large-scale Proteomics Survey. *J. Mol. Biol.* **362**, 915–924 (2006).

6. Appendices

Appendix 1. Remainder of clusters identified through Bagel not used for comparison

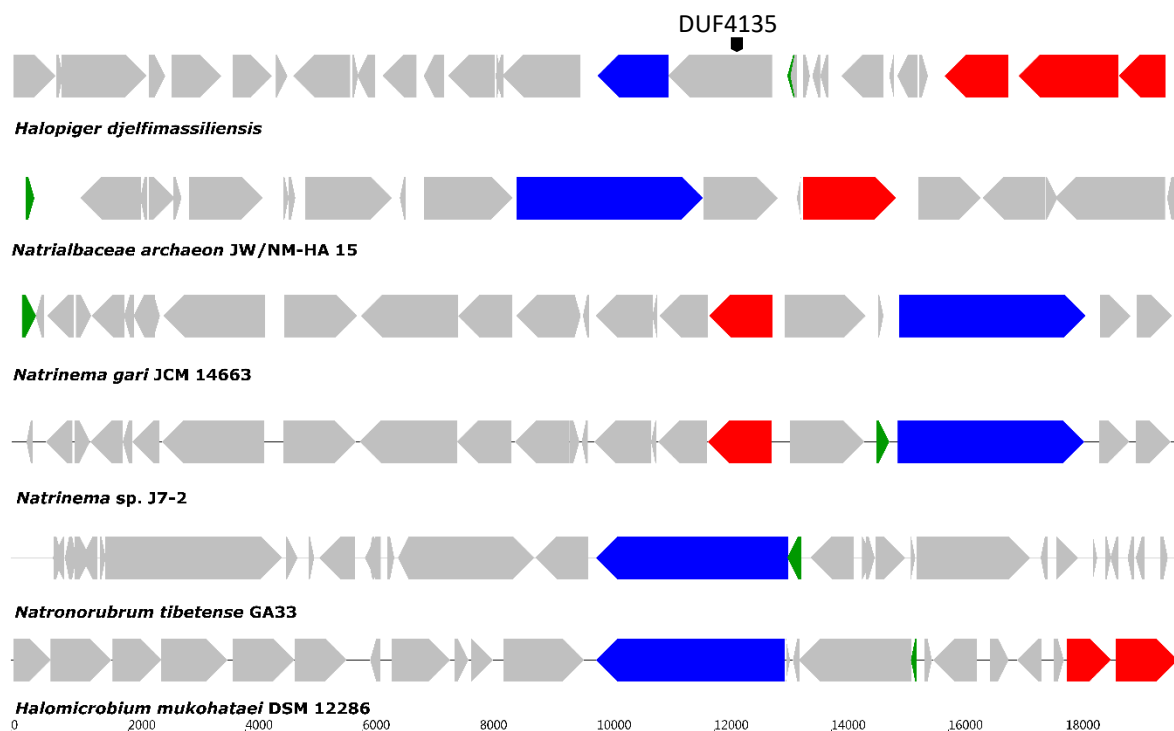


Figure 21 Putative lanthipeptide clusters identified through Bagel. In blue are the genes encoding the modifying enzymes LanM and in red the ABC transporter related gene. In green are the putative peptides identified by the software. Curiously, in the *Halopiger djelfimassiliensis* cluster the software identifies the domain DUF4135 (LanM dehydration domain) and the LanC domains as separate proteins. Homology between *Natrinema gari* JCM 14663 and *Natrinema sp. J7-2* might indicate that we are in the presence of the same organism.

Appendix 2. Genes present in chromosomal and plasmid clusters of *Haloferax mediterranei* ATCC 33500

Table 7 Genes present in chromosomal and plasmid clusters of *Haloferax mediterranei* ATCC 33500

Chromossome cluster	
HFX_RS04215	acetate--CoA ligase
HFX_RS04205	hypothetical protein
HFX_RS04200	methylmalonyl-CoA mutase
HFX_RS04195	Copper binding protein
HFX_RS04190	ABC transporter ATP-binding protein
HFX_RS04185	ABC transporter (transmembrane domain)
Plasmid cluster	
HFX_RS15975	TetR/AcrR family transcriptional regulator
HFX_RS15980	MFS transporter
HFX_RS15985	hypothetical protein
HFX_RS15995	oleate hydratase
HFX_RS16010	ABC transporter
HFX_RS16015	ABC transporter ATP-binding protein
HFX_RS16020	hypothetical protein

Appendix 3. Predicted masses for His₆-tagged LanA peptides for *E. coli* co-expression study

Table 8 Predicted masses for His₆-tagged LanA peptides for *E. coli* co-expression study. Each serine or threonine residue belonging to the core sequence can be subject to a dehydration reaction. Given that we have no means of ascertaining where leader sequence ends and core sequence begins an overestimation of the number of reactions is needed. For the loss of each water molecule, a shift of ~18Da is expected.

His ₆ -tagged LanA peptide	Number of dehydrations	Molecular mass in Da
RSFMedA1	No dehydration	6812,18
	1 dehydration	6794,18
	2 dehydrations	6776,18
	3 dehydrations	6758,18
	4 dehydrations	6740,18
	5 dehydrations	6722,18
	6 dehydrations	6704,18
	7 dehydrations	6686,18
	8 dehydrations	6668,18
RSFMedA2	No dehydration	9067,46
	1 dehydration	9067,46
	2 dehydrations	9067,46
	3 dehydrations	9067,46
	4 dehydrations	9067,46
	5 dehydrations	9067,46
	6 dehydrations	9067,46
RSFMedA3	No dehydration	8296,57
	1 dehydration	8278,57
	2 dehydrations	8260,57
	3 dehydrations	8242,57
	4 dehydrations	8224,57
	5 dehydrations	8206,57
	6 dehydrations	8188,57
	7 dehydrations	8170,57
	8 dehydrations	8152,57
	9 dehydrations	8134,57
	10 dehydrations	8116,57
	11 dehydrations	8098,57
	12 dehydrations	8080,57
	13 dehydrations	8062,57